Springer Proceedings in Mathematics & Statistics

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

Quantitative Psychology

83rd Annual Meeting of the Psychometric Society, New York, NY 2018

Springer

Springer Proceedings in Mathematics & Statistics

Volume 265

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

83rd Annual Meeting of the Psychometric Society, New York, NY 2018

Springer

Editors

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

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

Dylan Molenaar
Department of Psychology
University of Amsterdam
Amsterdam, The Netherlands

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

Jorge González
Facultad de Matematicas
Pontificia Universidad Catolica de Chile
Santiago, Chile

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-030-01309-7 ISBN 978-3-030-01310-3 (eBook)
https://doi.org/10.1007/978-3-030-01310-3

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

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

Preface

This volume represents presentations given at the 83rd annual meeting of the Psychometric Society, organized by Columbia University and held in New York, USA, during July 9–13, 2018. The meeting attracted 505 participants, and 286 papers were presented, of which 81 were part of a symposium. There were 106 poster presentations, 3 pre-conference workshops, 4 keynote presentations, 3 invited presentations, 2 career award presentations, 3 state-of-the-art presentations, 1 dissertation award winner, and 18 symposia.

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

We asked the 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 38 state-of-the-art chapters addressing a diverse set of psychometric topics, including item response theory, multistage adaptive testing, and cognitive diagnostic models.

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

Contents

Explanatory Item Response Theory Models: Impact on Validity and Test Development?

Susan Embretson

Abstract Many explanatory item response theory (IRT) models have been developed since Fischer's (Acta Psychologica 37:359–374, 1973) linear logistic test model was published. However, despite their applicability to typical test data, actual impact on test development and validation has been limited. The purpose of this chapter is to explicate the importance of explanatory IRT models in the context of a framework that interrelates the five aspects of validity (Embretson in Educ Meas Issues Pract 35, 6–22, 2016). In this framework, the *response processes* aspect of validity impacts other aspects. Studies on a fluid intelligence test are presented to illustrate the relevancy of explanatory IRT models to validity, as well as to test development.

Keywords Item response theory · Explanatory models · Validity

1 Introduction

Since Fischer (1973) introduced the linear logistic test model (LLTM), many additional explanatory IRT models have been developed to estimate the impact of item complexity on item parameters. These models include the linear partial credit model (LPCM; Fischer & Ponocny, 1995), the linear logistic test model with response error term (LLTM-R; Janssen, Schepers, & Peres, 2004), the constrained two parameter logistic model (2PL-Constrained; Embretson, 1999) and the Rasch facet model (Linacre, 1989). Explanatory IRT models also can include covariates for both items and persons, as well as within-person interactions (De Boeck & Wilson, 2004). Several models can detect strategy differences between persons, such as mixture distribution models (Rost, 1990; Rost & von Davier, 1995) and mixed models that include response time to detect strategies (Molenaar & De Boeck, 2018). Further, hierarchical models can be used in an explanatory fashion, such as item family models (Glas, van der Linden & Geerlings, 2010) and a criterion-referenced model (Janssen, Tuerlinckx, Meulder & De Boeck, 2000). Multidimensional IRT models with defined

S. Embretson (✉)
Georgia Institute of Technology, Atlanta, GA 30328, USA
e-mail: susan.embretson@psych.gatech.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_1

dimensions, such as the bifactor MIRT (Reise, 2012) or the multicomponent latent trait model (MLTM; Embretson, 1984, 1997) also can be used as explanatory IRT models. The *Handbook of Item Response Theory* (van der Linden, 2016) includes several explanatory models. Janssen (2016) notes that explanatory IRT models have been applied to many tests, ranging from mathematics, reading and reasoning to personality and emotions.

However, despite the existence of these models for several decades and their applicability to typical test data, actual impact on test development and validation has been limited. The purpose of this chapter is to highlight the importance of explanatory IRT models in test development. Studies on the development of a fluid intelligence test are presented to illustrate the use of explanatory IRT models in test design and validation. Prior to presenting the studies, background on the validity concept and a framework that unifies the various aspects are presented.

1.1 Test Validity Framework

In the current *Standards for Educational and Psychological Testing* (2014), validity is conceptualized as a single type (construct validity) with five aspects. First, the *content* aspect of construct validity is the representation of skills, knowledge and attributes on the test. It is supported by specified test content, such as blueprints that define item skills, knowledge or attribute representation, as well as specifications of test administration and scoring conditions. Second, the *response processes* aspect of validity consists of evidence on the cognitive activities engaged in by the examinees. These cognitive activities are assumed to be essential to the meaning of the construct measured by a test. The *Standards for Educational and Psychological Testing* describes several direct methods to observe examinees' processing on test items, such as eye-trackers movements, videos and concurrent and retrospective verbal reports/observations, as well as response times to items or the whole test. Third, the *internal structure* aspect of construct validity includes psychometric properties of a test as relevant to the intended construct. Thus, internal consistency reliability, test dimensionality and differential item functioning (DIF) are appropriate types of evidence. Item selection, as part of test design, has a direct impact on internal structure. Fourth, the r*elationship to other variables* aspect concerns how the test relates to other traits and criteria, as well as to examinee background variables (i.e., demographics, prior experience, etc.). Evidence relevant to this aspect should be consistent with the goals of measurement. Fifth, the *consequences* aspect of validity concerns how test use has adverse impact on different groups of examinees. While the test may not have significant DIF, studies may nonetheless show that the test has adverse impact if used for selection or placement. Adverse impact is particularly detrimental to test quality if based on construct-irrelevant aspects of performance.

The various aspects of validity can be conceptualized as a unified system with causal interrelationships (Embretson, 2017). Figure 1 organizes the five aspects into two general areas, internal and external, which concern test

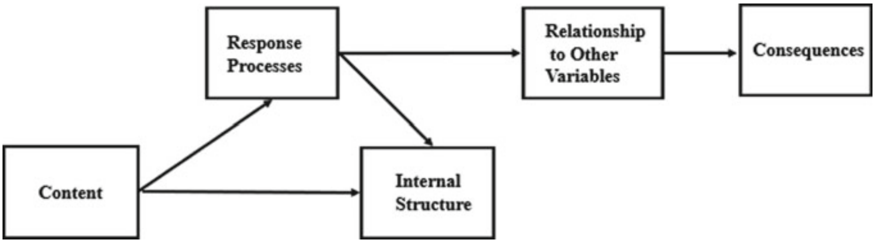


Fig. 1 Unified framework for validity

meaning and test significance, respectively. Thus, the *content*, *response processes* and *internal structure* aspects are relevant to defining the meaning of the construct while the *relationships to other variables* and *consequences* aspects define the significance of the test. Notice that the *content* and *response processes* aspect drive the other aspects causally in this framework. Importantly, these two aspects can be manipulated in test development. That is, item design, test specifications and testing conditions can impact test meaning. Thus, understanding the relationship between test content and response processes can be crucial in test development to measure the intended construct.

Unfortunately, the methods for understanding *response processes* described in the *Standards* have substantial limitations. Both eye-tracker data and talk aloud data are typically expensive to collect and analyze as well as impacting the nature of processing for examinees. Further, unless elaborated in the context of a model, the utility of response time data may be limited to identifying guessing or inappropriate responses. Importantly, explanatory IRT modeling can be applied to standard test data with no impact on examinees responses. Further, such models permit hypotheses to be tested about the nature of *response processes* through relationships of item content features and item responses.

2 Explanatory IRT Models in Item Design: Examples from ART

The Abstract Reasoning Test (ART) was developed in the context of research on *response processes*. ART is a test of fluid intelligence used to predict learning and performance in a variety of settings (Embretson, 2017). ART consists of matrix completion items as shown in Fig. 2. In these items, the examinee must identify the figure that completes the matrix based on the relationships between the figures across the rows and down the columns.

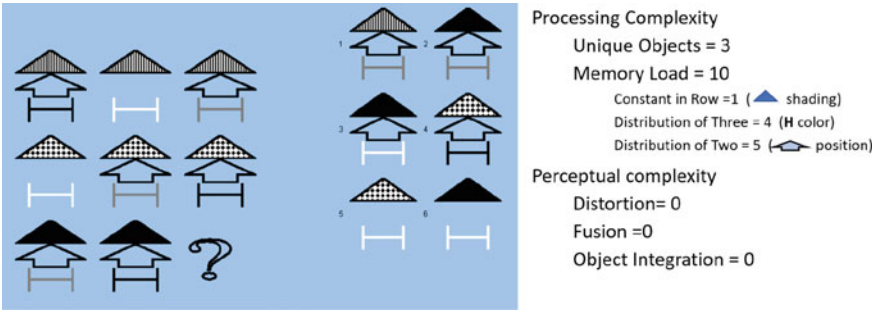


Fig. 2 Example of an ART item

2.1 Theory of Response Processes on Matrix Problems

Consistent with the Carpenter, Just and Shell's (1990) theory, it was hypothesized that examinees process the various elements individually in the matrix entries to find relationships. According to the theory, *processing complexity* is driven by the number of unique objects (as counted in the first entry) and memory load in finding relationships. Memory load depends on both the number and types of relationships, which are hypothesized to be ordered by complexity as follows: 1 = Constant in a Row (or column), the same figure appears in a row; 2 = Pairwise Progressions, figures change in the same way in each row; 3 = Figure Addition/Subtraction, the third column results from overlaying the first and second columns and subtracting common figures; 4 = Distribution of Three, a figure appears once and only once in each row and column and 5 = Distribution of Two, one figure is systematically missing in each row and column. Figure 2 illustrates relationships #1, #4 and #5 (see key on right) and Fig. 4 illustrates relationship #3. Relationship #2 could be illustrated by a change in object size across rows. Carpenter et al. (1990) postulate that these relationships are tried sequentially by examinees, such that Constant in a Row is considered before Pairwise Progressions and so forth. Thus, the Memory Load score is highest for the Distribution of Two relationships. Figure 2 shows numerical impact on Memory Load for three types of relationships. The difficulty of solving matrix problems also is hypothesized to depend on *perceptual complexity*, which is determined by Distortion, Fusion or Integration of objects in an entry. Figure 2 has none of these sources of *perceptual complexity* while Fig. 4 illustrates object integration in the matrix on the right side. Each matrix item can be scored for the *processing* and *perceptual complexity* variables. Item difficulty is postulated to result from these variables because they drive cognitive complexity.

2.2 Explanatory Modeling of Response Processes on ART Matrix Problems

An explanatory modeling of ART item difficulty results from applying LLTM to item response data, using the scores for matrix problem complexity. LLTM is given as follows:

$$P(\theta) = \frac{\exp(\theta_j - \sum_k \tau_k q_{ik} + \tau_0)}{1 + \exp(\theta_j - \sum_k \tau_k q_{ik} + \tau_0)} \quad (1)$$

where q_{ik} is the score for item *i* on attribute *k*, τ_k is the weight of attribute *k* in item difficulty and τ_0 is an intercept. Finally, θ_j is the ability of person *j*.

LLTM was applied to model item responses for ART items, scored for the two predictors of *processing complexity* and the three predictors of *perceptual complexity*. For example, a sample of 705 Air Force recruits were administered a form of ART with 30 items. The delta statistic, which is a likelihood ratio index of fit (Embretson, 1999) similar in magnitude to a multiple correlation, indicated that LLTM had strong fit to the data ($\Delta = .78$). The *processing complexity* variables had the strongest impact, especially memory load, which supports the theory.

2.3 Impact of Explanatory Modeling on Item Design for Matrix Problems

These results and the scoring system had direct impact on item and test design for ART. An automatic item generator was developed for ART items. Abstract structures were specified to define the objects within each cell of the 3 × 3 display and the response options. Types of relationships, as described above, specifies the changes in objects (e.g., circles, arrows, squares, etc.) and/or their properties (e.g., shading, borders, distortion, size, etc.) across columns and rows. LLTM results on military samples indicated high predictability of item difficulty by the generating structure (Δ = .90) and continued prediction by the five variables defining cognitive complexity ($\Delta = .79$).

3 Strategy Modeling in Test Design: Example from ART

Examinee differences in item solving strategies and potential impact on the various aspects of validity was examined in two studies. In Study 1, ART was administered with the original brief instructions. In Study 2, ART was administered with an expanded version of the instructions with examples of each type of relationship. In both studies, strategies were examined through mixture modeling.

3.1 Mixture Modeling to Identify Latent Classes

The mixture Rasch model (Rost & von Davier, 1995) can be applied to identify classes of examinees that vary in item difficulty ordering, which is postulated to arise from applying different item solving strategies. The mixture Rasch model is given as follows:

$$\mathrm{P}(\theta) = \Sigma_{\mathrm{g}}\pi_{\mathrm{g}}\frac{\exp(\theta_{\mathrm{jg}} - \beta_{\mathrm{ig}})}{1 + \exp(\theta_{\mathrm{jg}} - \beta_{\mathrm{ig}})} \qquad (2)$$

where β_{ig} is the difficulty of item *i* in class *g*, θ_{jg} is the ability of person *j* in class *g* and π_g is the probability of class *g*. Classes are identified empirically to maximize model fit. However, class interpretation can be examined by follow-up explanatory modeling (e.g., applying LLTM within classes) or by comparing external correlates of ability.

3.2 Study 1

Method. A form of ART with 30 items was administered to 803 Air Force recruits who were completing basic training. The ART instructions concerned the nature of matrix problems as defined by relationships in the row and columns in the 3 × 3 matrices. However, the scope of relationships that could be involved was not covered. ART was administered without time limits. Item parameters were estimated with the Rasch model and with the mixture Rasch model. In both cases the mean item parameter was set to zero.

Results from other tests were available on the examinees, including the *Armed Services Vocational Aptitude Battery* (ASVAB).

Results. The test had moderate difficulty for the sample based on raw scores ($M = 18.097$, $SD = 5.784$) and latent trait estimates ($M = .636$, $SD = 1.228$). Racial-ethnic comparisons were between groups with $N > 50$. The latent trait estimates were significant ($\mathrm{F}_{2,743} = 8.722$, $p < .001$, $\eta^2 = .023$). Standardized differences of ($d = .452$) for African Americans and ($d = .136$) for Hispanics were observed as compared to Caucasians.

The mixture Rasch model was applied with varying numbers of classes. Table 1 shows that while the log likelihood index (−2lnL) decreased successively from one to three classes, the Bayesian Information Criterion (BIC) increased for three classes. Thus, the two-class solution, with 68.7 and 31.2% of examinees in Class 1 and Class 2 respectively, was selected for further study. The latent trait means differed significantly between classes ($\mathrm{F}_{1,801} = 439.195, p < .001$), with Class 1 ($M = 1.143$, $SD = .984$) scoring higher than Class 2 ($M = -.413$, $SD = .865$). Significant racial ethnic differences were observed between the classes $\left(\chi^2_{1,695} = 12.958,\ p < .001\right)$, with 75.0% of Caucasians and 57.3% of African-Americans in Class 1.

Table 1 Mixture Rasch modeling results

Number of classes	Parameters	−2lnL	BIC
Study 1			
1	31	25,472	25,680
2	62	25,146	25,567
3	93	25,044	25,680
Study 2			
1	33	13,222	13,423
2	67	13,068	13,477
3	101	13,001	13,616

Table 2 LLTM weights, standard errors and t value by class

Complexity source	Class 1 (df = 572, Δ = .820)			Class 2 (df = 229, Δ = .809)		
	Weight	SE	t value	Weight	SE	t value
Unique elements	.1922	.0113	16.95*	.2681	.0185	14.50*
Memory load	.1851	.0049	37.49*	.0926	.0077	12.09*
Integration	.4543	.0454	10.00*	.5502	.0622	8.85*
Distortion	.7434	.0654	11.36*	−.0121	.1054	−.12
Fusion	.3150	.0508	6.20*	.0549	.0723	.76
Intercept	−4.1809	.1018	−41.08*	−2.2618	.1285	−17.61*

*p < .01

LLTM was applied within each class to determine the relative impact of the sources of cognitive complexity. While the overall prediction, as indicated by the Δ statistic (Embretson, 1999) shown on Table 2, was strong for both classes, the LLTM weights for cognitive complexity differed. Typically, the strongest predictor is Memory Load; however, the weight for Memory Load was significantly higher in Class 1. Unique Elements was the strongest predictor in Class 2 and two of three perceptual complexity variables were not significant.

Item difficulty also was modeled by the sources of memory load from the five types of relationships. It was found that the number of Figure-Addition relationships was correlated negatively for Class 1 ($r = -.211$) and positively for Class 2 ($r = .216$). Items with Figure-Addition relationships mostly more difficult for Class 2 (see Fig. 3).

Finally, ART trait estimates were correlated with four factors of ASVAB: Verbal, Quantitative, Perceptual Speed and Technical Information. Although significant positive correlations were found with all factors except Perceptual Speed for Class 1, no significant correlations with ASVAB factors were found for Class 2.

Discussion. Two classes of examinees, with varying patterns of item difficulty, were identified on the ART for fluid intelligence. Class 2 was characterized by substantially lower trait levels and lack of significant correlations with other aptitude

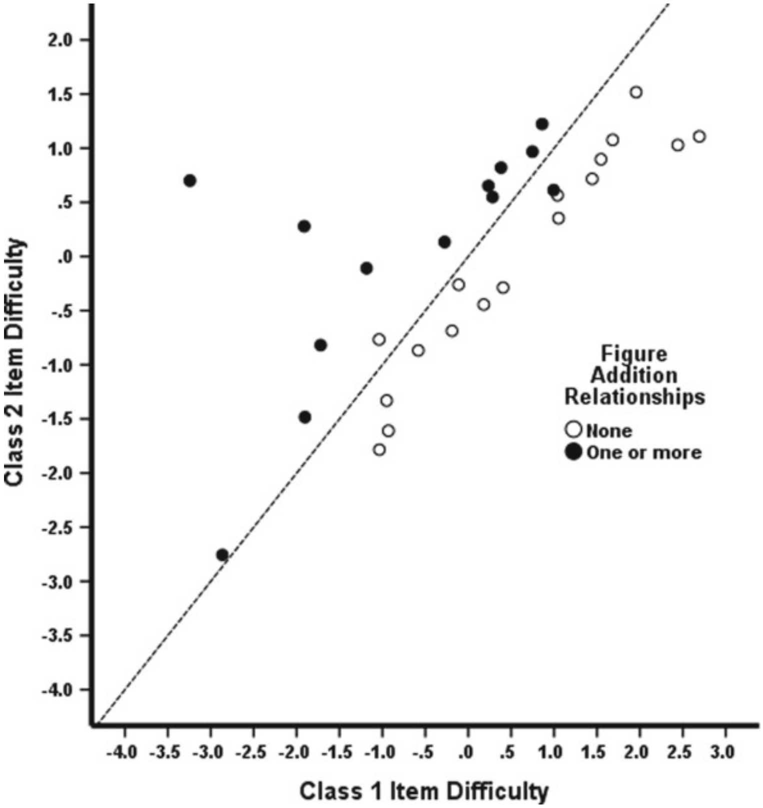


Fig. 3 Item difficulties by class

measures (i.e., ASVAB factors). Further, item difficulty was less predictable for Class 2 from the memory load associated with ART items. An analysis of the relationship types that contribute to memory load indicated that items with Figure-Addition relationships had substantially higher difficulty in Class 2. A possible explanation is that examinees in this class were unfamiliar with the Figure-Addition relationships and applied the much harder Distribution of Two relationship. Figure 4 shows examples of these relationships. Notice that the item on the left requires two Distribution of Two relationships (i.e., changes in the hourglass and house figures), as well as a Constant in a Row (triangles). The item on the right, however, can be solved by either three Figure-Addition (colum 3 is the substraction of column 2 from column 1) or three Distribution of Two relationships.

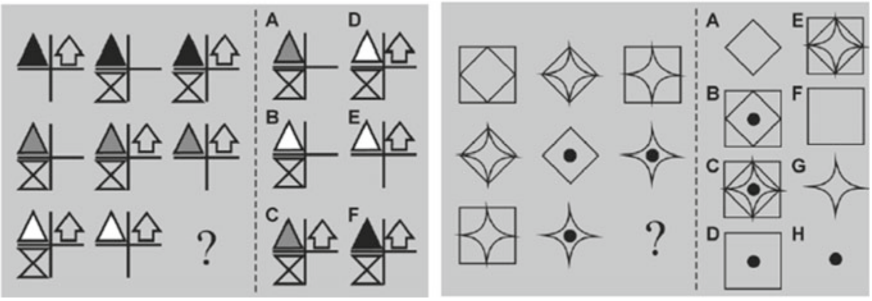


Fig. 4 Two ART items varying in distribution of two relationships

3.3 *Study 2*

The application of the mixture Rasch model in Study 1 identified a class of examinees with lower scores and different patterns of item difficulty that may be based on unfamiliarity with the possible types of relationships that can occur in ART. In this study, instructions were added to demonstrate each type of relationship.

Method. The examinees were 444 police recruits who were enrolled in basic training in law enforcement. A version of ART with 32 items included extended instructions in which all types of relationships were presented and illustrated. These instructions involved approximately eight additional minutes of testing time. For a sub-sample of examinees, training scores and scores on another test of fluid intelligence were available.

Results. The test was somewhat easy for the sample based on raw scores ($M = 21.459, SD = 4.779$) and latent trait estimates ($M = 1.152, SD = 1.203$). As for Study 1, racial-ethnic comparisons were made between groups with $N > 50$. The latent trait estimates were significant ($F_{2,406} = 3.099$, $p = .016$, $\eta^2 = .015$). Compared to Caucasians, standardized differences of ($d = .276$) for African Americans and ($d = .075$) for Hispanics were observed.

The mixture Rasch model was applied to determine the number of classes. Table 1 shows that while the log likelihood index ($-2\ln L$) decreased somewhat from one to two classes, the BIC index increased. Thus, the single class model is the preferred solution. Finally, for a subsample of 144 recruits, scores for a six-week course in legal issues for police officers were available. Training scores were correlated more highly with ART ($r = .333, p < .001$) than with the *Cattell Culture Fair Intelligence Test* (CCF; $r = .211, p = .009$).

Discussion. A single item-solving strategy is supported for ART when administered with extended instructions. That is, a single class was supported with mixture Rasch modeling. Further, the magnitude of the racial-ethnic differences was also substantially smaller in this study. Finally, ART correlated more highly with training than a similar non-verbal intelligence test, which has very short instructions.

4 Summary

The purpose of this chapter was to illustrate how using explanatory IRT models can contribute to the test development process and impact validity. The mixture Rasch model identified two classes of examinees on the ART with different item difficulty orders. The LLTM indicated strong predictability of item performance from cognitive complexity variables; however, the weights varied by class, supporting strategy differences. Items involving a certain type of relationship were relatively more difficult in the lower scoring class. Further, there was an undesirable impact of the second class on the *external relationships* aspect of validity; ART did not correlate with other aptitude tests and racial-ethnic differences were also found. A redesigned ART, that include extended instructions on types of relationships, had a single class, supporting common problem-solving strategies. Further, racial ethnic differences were substantially smaller on the redesigned ART and ART had stronger correlations with achievement than a similar test of fluid intelligence. Thus, two explanatory IRT models were used to inform the *responses processes* aspect of validity for a fluid intelligence test. The redesigned test to optimize responses processes had smaller racial-ethnic differences than the previous ART and more desirable external relationships than the CCF, a similar test of fluid intelligence.

References

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

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of processing in the Raven's Progressive Matrices Test. *Psychological Review, 97,* 404–431.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

Embretson, S. E. (1984). A general multicomponent latent trait model for response processes. *Psychometrika, 49,* 175–186.

Embretson, S. E. (1997). Multicomponent latent trait models. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305–322). New York: Springer.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64,* 407–433.

Embretson, S. E. (2016). Understanding examinees' responses to items: Implications for measurement. *Educational Measurement: Issues and Practice*, *35*, 6–22.

Embretson, S. E. (2017). An integrative framework for construct validity. In A. Rupp & J. Leighton (Eds.), *The handbook of cognition and assessment* (pp. 102–123). New York: Wiley-Blackwell.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.

Fischer, G. H., & Ponocny, I. (1995). Extended rating scale and partial credit models for assessing change. In G. H. Fischer & I. W. Molenaar (eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 353– 70). New York: Springer.

Glas, C. A. W., van der Linden, W. J., & Geerlings, H. (2010). Estimation of the parameters in an item cloning model for adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 289–314). New York: Springer.

Janssen, R. (2016). Linear logistic models. In W. van der Linden (Ed.), *Handbook of item response theory: Models, statistics and applications*. New York: Taylor & Francis Inc.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 189–212). New York: Springer.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion—referenced measurement. *Journal of Educational and Behavioral Statistics, 25,* 285–306.

Linacre, J. M. (1989). *Multi-facet Rasch measurement*. Chicago: MESA Press.

Molenaar, D., & De Boeck, P. (2018). Response mixture modeling: Accounting for heterogeneity in item characteristics across response times. *Psychometrika, 83,* 279–297.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47,* 667–696.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 3,* 271–282.

van der Linden, W. (Ed.). (2016). *Handbook of item response theory: Models, statistics and applications*. New York: Taylor & Francis Inc.

von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In G. Fischer & I. Molenaar (Eds.), *Rasch models: Foundation, recent developments and applications* (pp. 371–379). New York: Springer.

A Taxonomy of Item Response Models in Psychometrika

Seock-Ho Kim, Minho Kwak, Meina Bian, Zachary Feldberg, Travis Henry, Juyeon Lee, Ibrahim Burak Olmez, Yawei Shen, Yanyan Tan, Victoria Tanaka, Jue Wang, Jiajun Xu and Allan S. Cohen

Abstract The main aim of this study is to report on the frequency of which different item response theory models are employed in *Psychometrika* articles. Articles relevant to item response theory modeling in *Psychometrika* for 82 years (1936–2017) are sorted based on the classification framework by Thissen and Steinberg (Item response theory: Parameter estimation techniques. Dekker, New York, 1986). A sorting of the item response theory models used by authors of 367 research and review articles in Volumes 1–82 of *Psychometrika* indicates that the usual unidimensional parametric item response theory models for dichotomous items were employed in 51% of the articles. The usual unidimensional parametric item response theory models for polytomous items were employed in 21% of the articles. The multidimensional item response theory models were employed in 11% of the articles. Familiarity with each of more complicated item response theory models may gradually increase the percentage of accessible articles. Another classification based on recent articles is proposed and discussed. Guiding principles for the taxonomy are also discussed.

Keywords Item response theory · Models · Psychometrika · Rasch model · Taxonomy

1 Introduction

In this study, we report on the frequency of use of item response theory models in *Psychometrika* classified using the taxonomy of Thissen and Steinberg (1986) to answer the following questions: Will knowledge of a few basic item response theory models, such as the Rasch model and the three-parameter logistic model, assist readers in recognizing the modeling component of a high percentage of research articles that are relevant to item response theory modeling in *Psychometrika*? Which additional

S.-H. Kim (✉) · M. Kwak · M. Bian · Z. Feldberg · T. Henry · J. Lee · I. B. Olmez ·
Y. Shen · Y. Tan · V. Tanaka · J. Wang · J. Xu · A. S. Cohen
University of Georgia, Athens, GA 30602-7143, USA
e-mail: shkim@uga.edu
URL: https://coe.uga.edu/directory/people/shkim

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_2

item response theory models are used most often and therefore could be added most profitably to the psychometric and educational measurement background of readers? To aid psychometricians, measurement specialists, and applied statisticians who are continuing their own psychometric training, as well as persons designing courses in psychometrics and educational measurement for advanced undergraduate and graduate students, we report on modeling components of the item response theory relevant research and review articles in *Psychometrika* between 1936 and 2017.

In their taxonomy, Thissen and Steinberg (1986) classified item response theory models into four distinct groups based on assumptions and constraints on the parameters: binary models, difference models, divided-by-total models, and left-side-added models. They classified, for example, the two-parameter normal ogive model and the Rasch model as the binary models; Samejima's graded response model in normal ogive and logistic forms as the difference model; Bock's nominal response model and Master's partial credit model as the divide-by-total models; and Birnbaum's three-parameter logistic model as the left-side-added model (see Thissen & Steinberg, 1986, and references therein). In this paper, we present a more refined classification of the item response theory models based on the type of data analyzed.

2 Methods

This study analyzed Volumes 1 through 82 (March 1936 through December 2017) of *Psychometrika* and included all articles identified in the table of contents as Articles, Notes, Comments, Brief Comments, Tables, and Presidential Addresses. Excluded were Errata, Announcements, Abstracts, Book Reviews, Rules, Obituaries, Reports, Minutes, Notices, Constitution, and Lists of Members. For example, the excluded portions included: Volume 1, Issue 2, Pages 61–64 that contained the List of 150 Members of the Psychometric Society; Volume 4, Issue 1, Pages 81–88 that contained the List of 235 Members of the Psychometric Society; and Volume 2, Issue 1, Pages 67–72 that presented the Abstracts of 11 papers to be presented at the District Meeting of the Psychometric Society, The University of Chicago, on Saturday, April 3, 1987.

2.1 Review Process

Initially, 2837 articles were screened and identified from these volumes, and a group of measurement specialists eventually selected 367 articles for detailed review. The 367 articles were selected for their relevance to various models in item response theory. At least two measurement specialists independently reviewed each of the 367 articles for their use of item response theory models and completed a checklist documenting topics and models. The excluded articles received a second but briefer review for the presence or absence of the use of item response theory models in the procedures and techniques employed in the study. The reviewer read the abstracts, the methods sections, and all tables, and scanned other sections of the articles for the pertinent information. All reviewers were faculty members or graduate students trained in both quantitative methodology and applied statistics.

For the 367 articles receiving detailed review, any discrepancies between the two or more independent reviewers were discussed and resolved. Discrepancies were found initially for some of these articles. Another careful reading of these discrepant articles by the reviewers indicated that nearly all errors involved overlooking the methods section and the procedures and techniques used in the article.

For the 367 articles relevant to item response theory modeling in this study, we first partitioned these papers into theoretical and application types. Due to the characteristic of *Psychometrika* as a leading journal in psychometrics, the articles except for four were classified as theoretical.

2.2 Analysis of Models Used

We determined the frequency of the item response theory models in the 367 journal articles. Articles were sorted based on the classification framework by Thissen and Steinberg (1986). In addition to performing the simple quantification (number and percentage of articles using a method), we assessed how much a reader's acquaintance with additional item response theory models would improve his or her psychometric repertoire. In trying to obtain a definite measure, we were handicapped by the lack of a natural order for learning and applying these models. For the analysis, we chose the order that maximally increased the percentage of articles for which a reader would be acquainted with the item response theory models employed if he or she learned one more item response theory model.

We began this analysis by assuming that there are three major ordered classes of the item response theory models; (1) unidimensional parametric item response theory models for dichotomous items, (2) unidimensional parametric item response theory models for polytomous items, and (3) multidimensional item response theory models. In a sense, the order was thus determined by the complexity of models as well as modeling data gathered. This ordering, though useful, intellectually reasonable, and empirically based, is nevertheless arbitrary. In particular, it ignores the fundamental role of broad psychometric concepts used in the article such as adaptive testing, differential item functioning, equating and linking, parameter estimation techniques, test scoring, and so on in determining the extent of a reader's psychometric understanding. Furthermore, it may not be the best order for learning about the item response theory models.

3 Results

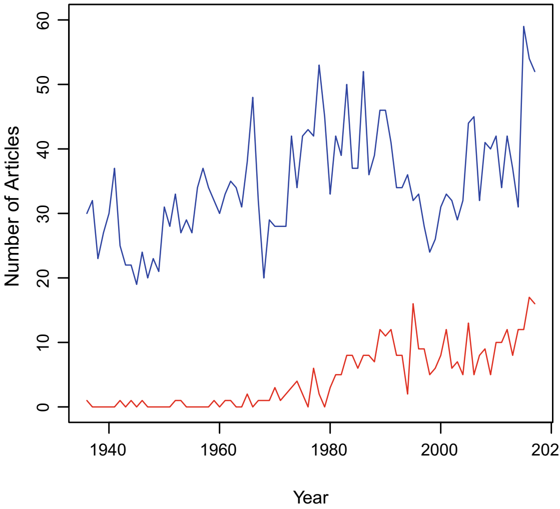
Figure 1 shows the time plots of the number of articles in *Psychometrika* as well as the number of item response theory relevant articles in each volume from 1936 to 2017. The average of the number of articles in each volume was 34.6 and its standard deviation was 8.6. The five number summary was (19, 28.8, 33, 41, 59). There was a steady increasing pattern in terms of the number of articles in each volume. The average of the number of item response theory relevant articles in each volume was 4.2 and its standard deviation was 4.3. The five number summary was (0, 0, 2.5, 8, 17). A rapid increase occurred between the 70's and the 90's for the number of item response theory relevant articles in each volume.

Figure 2 shows the time plot of the proportion of the item response theory relevant articles in each volume from 1936 to 2017. The average of the proportion was .11 and its standard deviation was .11. The five number summary was (0, 0, .07, .21, .53). The proportion was rapidly increased between the 70's and the 90's.

Table 1 presents the number of articles that used different item response theory models by decades from the 1930s (n.b., the 1930s starts from 1936) to the 2010s (n.b., the 2010s are not finished yet). The bottom line contains the total number of unique (i.e., not the column sum) item response theory relevant articles by decades. The far right-hand-side column of Table 1 shows the frequency of item response theory models found in Volumes 1 through 82 of *Psychometrika*. Under the assumptions outlined above, we analyzed the frequencies of the classes of item response theory models employed in the journal articles.

The followings are the model acronyms in Table 1: One-Parameter Logistic (1PL), One-Parameter Normal (1PN), Two-Parameter Logistic (2PL), Two-Parameter Normal (2PN), Nonparametric (NON), Three-Parameter Logistic (3PL), Three-Parameter

Fig. 1 Time plots of the number of articles in blue and the number of item response theory relevant articles in red



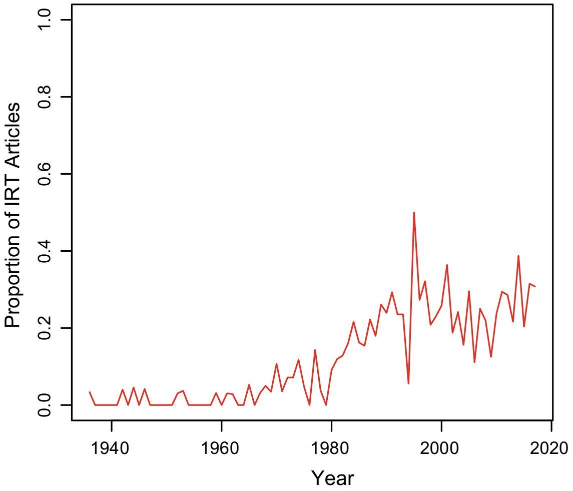


Fig. 2 Time plot of the proportion of item response theory relevant articles

Normal (3PN), Two-Parameter of Choppin (2PC), Four-Parameter Logistic (4PL), Multiple Choice of Samejima (MCS), Multiple Choice of Thissen and Steinberg (MCTS), Multiple Choice (MC), Graded Response (GR), Partical Credit (PC), Rating Scale (RS), Generalized Partial Credit (GPC), Nominal Categories (NC), Binomial Trials (BT), Poisson Counts (POC), Continuation Ration (CR), Linear Logistic Test Model (LLTM), and Multidimensional Item Response Theory (MIRT).

Table 1 shows many articles reviewed relied on some type of unidimensional dichotomous item response theory models. These articles used the Rasch model most frequently by 107 out of 367 articles. The one-parameter logistic model with a common item discrimination parameter was used in 15 articles. The two-parameter logistic model was used by 60 out of 367 articles, and the two-parameter normal ogive model was used by 37 out of 367 articles. The three-parameter logistic model was used quite frequently, that is, 82 out of 367 articles. The polytomous item response theory models are generally used less frequently (25 for the graded response model, 21 for the partial credit model, 10 for the rating scale model, 5 for the generalized partial credit model, and 7 for the nominal categories model).

It can be noticed that the various taxonomic classifications of the item response theory models defined in Table 1 were not frequently employed in the articles reviewed. The impression is that only limited cases of the item response theory models or the combinations of the models have been employed in *Psychometrika*, although this finding does depend on the initial taxonomy of the item response theory models. Articles published recently within about 20 years that used item response theory models were more complicated both mathematically and statistically than other previously published articles in *Psychometrika*. Theoretical research studies based on more complicated item response theory models require a deeper understanding of and more extensive training in psychometrics and applied statistics.

Table 1 Item response theory models from psychometrica articles

Taxonomy type	Model	Period									
		1930s	40s	50s	60s	70s	80s	90s	2000s	10s	Row total
Binary	Rasch					7	33	27	15	25	107
	1PL						6	2	1	6	15
	1PN										
	2PL			1	2		11	16	9	21	60
	2PN	1	3	3	5	2	5	4	8	6	37
	NON						1	7	9	6	23
Left-side-added (LSA)	3PL					6	20	19	14	23	82
	3PN						1				1
	2PC										
	4PL								1	1	2
LSA-DBT	MCS										
	Model 6										
	MCTS						1				1
	MC								1	1	2
Difference	GR				2	4	2	9	2	6	25
Divided-by-total (DBT)	PC						5	7	3	6	21
	RS					1	5	1	2	2	10
	GPC							4	1		5
	NC					2	1	1	3		7
	BT										
	POC							2	1		3
	CR					2			1		3
Extension	LLTM						3	2	6	1	12
	MIRT						1	6	10	17	34
	Testlet							1		1	2
	Multilevel								3	3	6
	Other						5	4	9	5	23
Unique articles	No.	1	3	3	7	23	70	85	78	97	367

A sorting of the item response theory models used by authors of the 367 articles in *Psychometrika* indicates that a reader who is familiar with the usual unidimensional parametric item response theory models for dichotomous items (e.g., the Rasch model, the one-parameter logistic model, the two-parameter logistic or normal ogive model, and the three-parameter logistic or normal ogive model) may have potential access to 186 out of 367 articles (51%). Note that the number 186 was not obtained from Table 1 but based on the separate counting of the articles. Note also that the numbers in Table 1 are not mutually exclusive because, for example, an article might employ two or more different item response theory models together. It should also be noted that the accessibility here implies the recognition of the model used in the article instead of comprehension of the entire contents of the article. Because the unidimensional parametric item response theory models for polytomous items (e.g., the graded response model, the partial credit model, the rating scale model, the nominal categories model, and the generalized partial credit model) were employed in 79 out of 367 articles, a reader who is familiar with the two classes of the unidimensional item response theory models may have potential access to cumulatively 72% of the journal articles. Familiarity with each of the more complicated item response theory models may gradually increase the percentage of accessible articles. If one knew the multidimensional item response theory models in addition to the unidimensional item response theory models, one would access 38 articles, or 83 cumulative per cent of the number of articles reviewed. However, more complicated models (e.g., nonparametric models, testlet models, mixture models, multilevel models, etc.) were concurrently used in the psychometric research journal articles together with the usual parametric models for the dichotomous and polytomous items. Hence, 64 out of 367 (17%) of the articles cannot be fully accessible in terms of item response theory if a reader is familiar with only these parametric models.

Although some classifications were obviously quite narrowly defined, others such as multidimensional item response theory models and nonparametric models were not. Furthermore, these latter models, though cited infrequently in the articles, may be more frequently used in other application fields and may become more common in future psychometric research.

The selected articles relevant to item response theory modeling in Table 1 were sorted based on the classification framework by Thissen and Steinberg (1986). Another recent classification based on Van der Linden (2016a), however, can be used, and a more refined subclassification (e.g., Nering & Ostini, 2010) can also be considered. Note that articles may be further sorted by the parameter estimation methods (e.g., Baker & Kim, 2004; De Ayala, 2009) as well as the computer programs used to implement the estimation methods (e.g., Hambleton, Swaminathan, & Roger, 1991, pp. 159–160; Van der Linden, 2016b).

Psychometric researchers interested in continuing their own training in methodology should find the frequencies of various item response theory models presented in Table 1 helpful in identifying the knowledge of which item response theory models they should be aware. This paper reviews item response theory models with the perspective of a general reader, and no attempt has been made to identify a hierarchical structure of the item response theory models, which may vary for researchers in different psychometric research areas within different specialties.

4 Discussion

Except for general item response theory review articles in *Psychometrika*, not many item response theory models were used simultaneously in each research article. As noted by Popham (1993) and Bock (1997) there are several unexpected consequences of using item response theory models in the analysis of testing data. Only a limited number of item response theory experts can fully understand what is happening, for example, in the process of test calibration. Also, there are many different directions of the development of item response theory so that even experts in the item response theory field may not be able to comprehend the full scope of the theory and applications of item response theory. It is unfortunate that the item response theory models and item response theory itself are too difficult to understand for scholars with only limited statistical and mathematical training. Nevertheless, item response theory does occupy and may continue to occupy major portions of lively and productive future development in psychometric research.

Understanding some of the item response theory relevant articles in *Psychometrika* definitely requires more than the familiarity of the item response theory models. For example, training in modern Bayesian statistics for which the Bayesian posterior approximation methods with data augmentation techniques are taught is needed for reading several articles. Note that the normal ogive models were frequently employed in data augmentation techniques by some authors who are themselves prepared for understanding more advanced research articles.

It should be noted that the numerical measure of ability or proficiency is the ultimate, eventual entity that is pursued in item response theory modeling. In other applications, the item parameters are something needed assuming that persons are randomly sampled from a population. In item response theory with such a sampling concept, the item parameters are the structural parameters while the person parameters are incidental parameters. The concept of invariance of ability with regard to the sets of item parameters (i.e., persons ability can be measured with different sets of items) as well as invariance of item characteristics with regard to the groups of persons (i.e., item characteristics can be obtained with different groups of persons) are crucial in item response theory. Many investigations of structural parameters such as measurement invariance or differential item functioning studies are studies of structural parameters. Note that measurement invariance is a preliminary to studying invariant person measures, and as such, needs to be seen as a process within measurement validation (Kane, 2006). Hence, item response theory models and the required parameters to estimate ought to be scrutinized in conjunction with specific application areas.

In the field of educational assessment, items can be in the forms of both dichotomously-scored and polytomously-scored. In most large scale assessment programs (e.g., National Assessment of Educational Progress, Trends in International Mathematics and Science Study) a combination of the three-parameter logistic model and the generalized partial credit model is used to calibrate item response data. In the analysis of instruments with mixed item types, there are special combinations of dichotomous and polytomous item response theory models to be used (e.g., Rasch and PC; 2PL and GR). So, there are natural combinations of item response theory models for mixed item types.

This study may be helpful to people designing and teaching courses in psychometric methods for advance undergraduate and graduate students and other psychometricians or measurement specialists using various item response theory models. But one should keep in mind that any professional specialization in psychometric research may influence understanding with regard to the relative importance of the various item response theory models.

The purpose of writing for some journal articles that are relevant to item response theory in psychometric research might not be to disseminate the findings of the studies to more general psychometric researchers. The authors might have tried to demonstrate their capabilities to invent novel models, to create new ideas, and to explore challenging areas of psychometrics. Consequently, there are a plethora of item response theory models invented recently.

We have identified the various models in item response theory that have been used by psychometricians in *Psychometrika* articles and that are thus very much likely to be used by future authors in psychometric research. Note that the latter point may not be the case because some articles used the most esoteric item response theory models together with complicated computational techniques. The appropriate training of psychometric researchers in the use of item response theory models seems to be an important consideration. Such an issue should be addressed by the leading scholars who are responsible for training future psychometric researchers. More in depth evaluation of the articles and more thorough review would be helpful.

It can be noted that item response theory models presented in Table 1 already contained additional models than those (e.g., 4PL, MC, CR, LLTM, MIRT, Testlet, and Multilevel) in Thissen and Steinberg (1986). There are many different item formats so we may classify item response theory models in terms of the item response data and additional variables required for the modeling. If we denote the original item response data for multiple-choice items as U, then item response theory models for multiple-choice items can be used to estimate model parameters. When we denote the keyed or scored data to be R and further denote dichotomously scored data to be D, then we may use the Rasch model and other item response theory models for dichotomously scored items (e.g., 1PL, . . . , 4PL). If R can be further specified with the types of polytomously scored items that is denoted by P, then we may use item response theory models for polytomous items. Here, the set of item parameters can be denoted by ξ and the set of ability parameters can be denoted by θ. If we allow additional dimensionality to the item and ability parameters, then we may have multidimensional item response theory models. In the above context, if there exist a latent group hyperparameter τ and both ability and item parameters are characterized by τ, then we may have mixture item response theory models. Note that in all of these models, we are not required to use any auxiliary variables.

When examinee groups are organized with a manifested variable g (e.g., male and female), then the data can be seen as Rg. In conjunction with item response theory modeling, such data can be denoted as the multiple group data (e.g., differential item functioning data, equating and linking data, measurement invariance data, etc.). Models for differential item functioning, for example, can be applied to such data. A similar case with a timing variable t can yield data Rt. The item response theory model for such data can be denoted as the parameter drift model. It should be noted that in case of obtaining R, raters can be entered to the modeling as a new facet of the resulting data in the generalizability context. If the raters' information denoted as r is also entered to the data, then we may express the data Rr and add the set of raters' severity parameters to the model. If there exists information about the cognitive components and the items (e.g., W for the linear logistic test model), the resulting models relate ξ to another set of basic parameters (e.g., η). There are several linear logistic or component test models that use RW as input data. The testlet model seems to require an additional vector that contain item relationship or dependency d, that is, Rd as the input data and tries to explicate the dependency among the items. If there are the matrix of nested grouping structure G added to the data, then the required item response theory model for analyzing such data of RG becomes the multilevel item response theory model. In addition, the time matrix T that contains examinees' response time to items is analyzed that yields RT, then item response theory models that contain speededness parameters may be used.

In the above classification, the type of input data determines the resulting so called parametric item response theory models. There are also nonparametric item response theory models for these data as well as models for nonmonotone items. The item response theory models, hence, can be classified with the input data type as well as the characteristics of item response functions and the methods of parameter estimation. Note that Thissen and Steinberg's (1986) classification was partly based on the parameter estimation context.

References

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Dekker.

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice*, *16*(4), 21–32.

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

Hambleton, R. K., Swaminathan, H., & Roger, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (6th ed.). Westport, CT: Praeger.

Nering, M. L., & Ostini, R. (2010). *Handbook of polytomous item response theory models*. New York, NY: Routledge.

Popham, W. J. (1993). Educational testing in America: Whats right, whats wrong? A criterion referenced perspective. *Educational Measurement: Issues and Practice*, *12*(1), 11–14.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567–577.

Van der Linden, W. J. (Ed.). (2016a). *Handbook of item response theory, Volume 1: Models.* Boca Raton, FL: CRC Press.

Van der Linden, W. J. (Ed.). (2016b). *Handbook of item response theory, Volume 3: Applications.* Boca Raton, FL: CRC Press.

NUTS for Mixture IRT Models

Rehab Al Hakmani and Yanyan Sheng

Abstract The No-U-Turn Sampler (NUTS) is a relatively new Markov chain Monte Carlo (MCMC) algorithm that avoids the random walk behavior that common MCMC algorithms such as Gibbs sampling or Metropolis Hastings usually exhibit. Given the fact that NUTS can efficiently explore the entire space of the target distribution, the sampler converges to high-dimensional target distributions more quickly than other MCMC algorithms and is hence less computational expensive. The focus of this study is on applying NUTS to one of the complex IRT models, specifically the two-parameter mixture IRT (Mix2PL) model, and further to examine its performance in estimating model parameters when sample size, test length, and number of latent classes are manipulated. The results indicate that overall, NUTS performs well in recovering model parameters. However, the recovery of the class membership of individual persons is not satisfactory for the three-class conditions. Findings from this investigation provide empirical evidence on the performance of NUTS in fitting Mix2PL models and suggest that researchers and practitioners in educational and psychological measurement should benefit from using NUTS in estimating parameters of complex IRT models.

Keywords Markov chain Monte Carlo · No-U-Turn sampler · Mixture IRT models

1 Introduction

Classical test theory (CTT; Novick, 1966) has served the measurement community well for most of the last century. However, problems emerged using CTT have encouraged the development of a modern test theory, namely the item response theory (IRT; Lord, 1980), which has become a fundamental tool for measurement professionals

R. Al Hakmani (✉) · Y. Sheng
Southern Illinois University, 62901 Carbondale, IL, USA
e-mail: rehab.hekmani@siu.edu

Y. Sheng
e-mail: ysheng@siu.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_3

in behavioral sciences (van der Linden & Hambleton, 1997). IRT consists of a family of models that specify the probability of a response given person latent trait and item characteristics. Different models exist for different types of response data. Conventional dichotomous IRT models (e.g., Birnbaum, 1969; Lord, 1980; Lord & Novick, 1968; Rasch, 1960), including the one-parameter logistic (1PL), the two-parameter logistic (2PL), and the three-parameter logistic (3PL) models, are used when test items require binary responses such as true-false questions or multiple-choice questions that are scored as correct or incorrect.

These conventional IRT models assume that the observed response data stem from a homogenous population of individuals. This assumption, however, limits their applications in test situations where, for example, a set of test items can be solved with different cognitive strategies. If the population consists of multiple groups of persons, with each group employing a different strategy for the same item, the parameters for this item will be different across these groups (or subpopulations), and consequently, the conventional IRT models cannot be used for the response data. On the other hand, the conventional IRT models may hold when each of the subpopulations employs a common strategy. As a result, mixture IRT (MixIRT; Rost, 1990) models have been developed to capture the presence of these latent classes (i.e. latent subpopulations) that are qualitatively different but within which a conventional IRT model holds. MixIRT models have become increasingly popular as a technique for investigating various issues in educational and psychological measurement such as identifying items that function differently across latent groups (e.g., Choi, Alexeev & Cohen, 2015; Cohen & Bolt, 2005; De Ayala, Kim, Stapleton, & Dayton 2002; Maij-de Meij, Kelderman, & van der Flier, 2008; Samuelsen, 2005; Shea, 2013; Wu et al., 2017) or detecting test speededness (e.g., Bolt, Cohen, & Wollack, 2002; Meyer, 2010; Mroch, Bolt, & Wollack, 2005; Wollack, Cohen, & Wells, 2003).

Over the past decades, the estimation of IRT and particularly MixIRT models has moved from the traditional maximum likelihood (ML) approach to the fully Bayesian approach via the use of Markov Chain Monte Carlo (MCMC) techniques, whose advantages over ML have been well documented in the IRT literature (e.g., de la Torre, Stark, & Chernyshenko, 2006; Finch & French, 2012; Kim, 2007; Wollack, Bolt, Cohen, & Lee, 2002). The common MCMC algorithms, such as Gibbs sampling (Geman & Geman, 1984) and Metropolis Hastings (MH; Hastings, 1970; Metropolis & Ulam, 1949), have been applied to estimate MixIRT models (e.g., Cho, Cohen, & Kim, 2013; Huang, 2016; Samuelsen, 2005; Shea, 2013). These algorithms, however, suffer from problems of inefficiently exploring the parameter space due to their random walk behavior (Neal, 1992). Recent developments of MCMC focus on non-random walk MCMCs such as the no-U-turn sampler (NUTS; Hoffman & Gelman, 2011), which can converge to high dimensional posterior distributions more quickly than common random walk MCMC algorithms, and is hence less computational expensive. In the IRT literature, Zhu, Robinson, and Torenvlied (2015) applied NUTS to simple IRT models and demonstrated its advantage over Gibbs sampling in the efficiency of the algorithm. Although NUTS has been applied with simple unidimensional IRT models (e.g., Chang, 2017; Luo & Jiao, 2017; Grant, Furr, Carpenter,

& Gelman, 2016), to date, no research has investigated its application to the more complex IRT models, such as MixIRT models.

1.1 Two-Parameter Mixture IRT Model

In the MixIRT modeling framework, persons are characterized by their location on a continuous latent dimension as well as by their latent class membership. Also, each subpopulation has a unique set of item parameters (e.g., difficulty, or discrimination). This study focuses on the two-parameter mixture (Mix2PL) IRT model, which can be viewed as an extension of the mixture Rasch model proposed by Rost (1990). If we let Y_{ij} detonate a correct ($Y_{ij} = 1$) or incorrect ($Y_{ij} = 0$) response for person i to item j, the probability of a correct response in the Mix2PL model is defined as

$$\begin{aligned} P\left(Y_{ij} = 1|\theta\right) &= \sum_{g=1}^{G} \pi_g \times P\left(Y_{ij} = 1\middle|\theta_{ig}, b_{jg}, a_{jg}, g\right) \\ &= \sum_{g=1}^{G} \pi_g \times \frac{\exp\left[a_{jg}\left(\theta_{ig} - b_{jg}\right)\right]}{1 + \exp\left[a_{jg}\left(\theta_{ig} - b_{jg}\right)\right]}, \end{aligned} \tag{1}$$

where $g = 1, \ldots, G$ is the latent class indicator, θ_{ig} denotes the ability for person i in class g, π_g denotes the proportion of examinees (i.e., the mixing proportion) in each class with a constraint that all these proportions sum to one, and b_{jg} and a_{jg} are the difficulty and discrimination parameters, respectively, for item j in the gth class.

1.2 Non-random Walk MCMC

Random walk algorithms such as Gibbs sampling and MH explore the parameter space via inefficient random walks (Neal, 1992). For complicated models with many parameters, these methods may require an unacceptably long time to converge to the target posterior distribution. On the other hand, non-random walk algorithms such as Hamiltonian Monte Carlo (HMC; Duane, Kennedy, Pendleton, & Roweth, 1987) and NUTS avoid the inefficient exploration of the parameter space. Specifically, HMC borrowed its idea from physics to suppress the random walk behavior by means of an auxiliary variable, momentum, that transforms the problem of sampling from a target posterior distribution into the problem of simulating Hamiltonian dynamics, allowing it to move much more rapidly through the posterior distribution (Neal, 2011). The unknown parameter vector $\boldsymbol{\theta}$ is interpreted as the position of a fictional particle. The Hamiltonian is an energy function for the joint state of the position $\boldsymbol{\theta}$ and the momentum $\boldsymbol{\phi}$, which defines a joint posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{y})$. At each iteration, a random momentum vector $\boldsymbol{\phi}$ is generated, which is usually drawn from a multivariate normal distribution $N(\mu, \boldsymbol{\Sigma})$ with mean μ and covariance matrix $\boldsymbol{\Sigma}$. Then, the path of the particle is simulated with a potential energy equal to the negative value of the log of the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$.Values of $(\boldsymbol{\theta}, \boldsymbol{\phi})$ are simultaneously updated over time using the leapfrog algorithm, which breaks the time into discrete steps such that the total Hamiltonian simulation time is the product of the discretization interval (or the step size ε) and the number of steps taken per iteration (or the leapfrog steps L). After a Metropolis decision step is applied, the whole process repeats for an adequate number of iterations until convergence is reached (Gelman, Carlin, Stern, Dunson, Vehtari, & Rubin, 2014; Stan Development Team, 2017).

Although HMC is a powerful MCMC technique, it requires choosing suitable values for three parameters (i.e., the step size ε, the number of leapfrog steps L, and the mass matrix $\boldsymbol{\Sigma}$) for the fictional particle. Tuning these parameters, and specifically L, requires expertise and a few preliminary runs, which can be challenging (Neal, 2011; Hoffman & Gelman, 2011). To overcome this, Hoffman and Gelman (2011) introduced NUTS to eliminate the need to set the number of leapfrog steps that the algorithm takes to generate a proposal state. Using a recursive algorithm, NUTS creates a set of candidate points that spans a wide path of the target posterior distribution, stopping automatically when it starts to double back and retrace its steps (i.e. starts to make a U-turn). Empirically, NUTS performs as efficiently as, and sometimes better than, a well-tuned HMC without requiring user interventions. Thus, NUTS is a tune-free technique, which will make it easily accessible by practitioners and researchers in behavioral sciences to fit various complex measurement models.

In view of the above, the purpose of this study is to investigate how NUTS performs in recovering parameters of the Mix2PL model under various test conditions where sample size, test length, and number of latent classes are taken into consideration. The significance of the study lies in that it not only demonstrates the application of a more efficient MCMC algorithm to the more complex MixIRT model, but also provides guidelines to researchers and practitioners on the use of such models under the fully Bayesian framework. The successful implementation of NUTS to the Mix2PL model will also help researchers with fitting more complex IRT models using fully Bayesian estimation. Findings from this investigation will provide empirical evidence and shed light on the performance of NUTS in fitting more complicated IRT models.

1.3 Model Identification

Given the difference between Bayesian and likelihood identifiability (Gelfand & Sahu, 1999), the Mix2PL IRT model was identified under the fully Bayesian approach following the literature to avoid two problems: (a) the indeterminacy and (b) the problem of label switching, which is inherent in mixture models in general. The usual practice to avoid the indeterminacy in MixIRT models, as recommended by Rost (1990), is to impose a sum-to-zero constraint in the item difficulty parameter within each latent class (i.e., $\sum_j b_{jg} = 0$). Under the fully Bayesian estimation using NUTS, there are several methods available to enforce a sum-to-zero constraint on a parameter vector (see Stan Development Team, 2017 for more details). Due to its ease in implementation, soft centering was used in this study to apply the sum-to-zero constraint on the difficulty parameter in each latent class (i.e., $b_g \sim N(0, 1)$). Further, one practice for avoiding the problem of label switching of mixture components in MixIRT models, under the fully Bayesian framework, is to impose an ordinal constraint on the mixing proportions or the difficulty parameter (e.g., Bolt et al., 2002) or other parameters (e.g., Meyer, 2010) across latent classes. In this study, an ordinal constraint had to be imposed on both the mean ability (μ_g) parameters and the item difficulty parameters (b_g) to ensure Bayesian identifiability with Mix2PL models.

2 Methods

Monte Carlo simulations were carried out to investigate the performance of NUTS in terms of parameter recovery of the Mix2PL model under various test conditions. Data were generated using the Mix2PL model as defined in Eq. (1) with equal proportions (i.e., equal class sizes) while manipulating three factors: test length ($J = 20$ or 30), number of latent classes ($G = 2$ or 3), sample size in each subpopulation ($n = 250$ or 500). Specifically, for the two-class condition ($G = 2$), the total number of subjects (N) was 500 or 1000; the mixing proportions were $\pi_1 = 0.50$ and $\pi_2 = 0.50$; the person ability parameters were generated from a mixture of two subpopulations where $\theta_1 \sim N(-2, 1)$ and $\theta_2 \sim N(2, 1)$; the class-specific item difficulty parameters were generated from a uniform distribution where $b_1 \sim U(-2, 0)$ and $b_2 \sim U(0, 2)$; and the class-specific item discrimination parameters were generated from a uniform distribution where $a_g \sim U(0, 2)$, $g = 1$ or 2. For the three-class condition ($G = 3$), the total number of subjects was 750 or 1500; the mixing proportions were $\pi_1 = 0.33$, $\pi_2 = 0.33$, and $\pi_3 = 0.33$; the person ability parameters were generated from a mixture of three subpopulations where $\theta_1 \sim N(-4, 1)$, $\theta_2 \sim N(0, 1)$, and $\theta_3 \sim N(4, 1)$; the class-specific item difficulty parameters were generated from a uniform distribution where $b_1 \sim U(-2, -0.5)$, $b_2 \sim U(-0.5, 0.5)$, and $b_3 \sim U(0.5, 2)$; and the class-specific item discrimination parameters were generated from a uniform distribution where $a_g \sim U(0, 2)$, $g = 1, 2$, or 3.

Priors and hyperpriors were selected to be comparable to those adopted by others (e.g., Bolt, Cohen, & Wollak, 2002; Meyer, 2010; Li, Cohen, Kim, & Cho, 2009; Wollack et al., 2003). Specifically, normal prior densities were used for person ability parameters $\theta_{ig} \sim N(\mu_g, 1)$, with a standard normal distribution for the hyperparameters μ_g, and a Dirichlet distribution for the mixing-proportion parameters such that $(\pi_1, \ldots, \pi_G) \sim \text{Dirichlet}(1, \ldots, 1)$.

Convergence of the Markov chains was examined using the Gelman-Rubin R statistic (Gelman & Rubin, 1992), with a threshold of 1.10 as suggested by Gelman et al. (2014). For the conditions involving two latent classes, the warm-up stage of either 2000 or 3000 iterations followed by 3 chains with either 3000 or 5000 sampling iterations was sufficient for the chains to reach convergence when the sample size was 500 or 1000, respectively. For the conditions involving three latent classes, in order to reach convergence, the warm-up stage had to reach 3000, 5000 or 8000 iterations followed by 3 chains with 5000, 7000 or 10,000 sampling iterations for $N = 750$ or $N = 1000$, respectively. Ten replications were conducted for each of the simulated condition. The precision of the class and item parameter estimates was evaluated using bias and root mean square error (*RMSE*), which are defined as

$$bias_{\xi} = \frac{\sum_{r=1}^{R}\left(\widehat{\xi_r} - \xi\right)}{R}, \tag{2}$$

$$RMSE_{\xi} = \sqrt{\frac{\sum_{r=1}^{R}\left(\widehat{\xi_r} - \xi\right)^2}{R}}, \tag{3}$$

where ξ is the true value of the parameter (e.g., μ_g, π_g, a_{jg}, or b_{jg}), and ˆξ is the estimated value of the parameter in the *r*th replication where $r = 1, \ldots, R$. To summarize the recovery of item parameters, these measures were averaged over items. Further, the recovery of class memberships was evaluated by computing the percentage of correct classifications of individual persons into the class from which they were simulated. This was achieved by first calculating the probability of membership in each class *g* for each individual. Then, each individual was assigned to the latent class for which he or she has the highest probability of belonging (i.e., the largest membership probability).

3 Results

3.1 Mixing-Proportion and Mean Ability Recovery

The results for recovering the mixing proportion and mean ability for each latent class in the Mix2PL model are summarized in Tables 1 and 2 for the two- and three-class conditions, respectively. The small values of *bias* and *RMSE* suggest that NUTS performed well in recovering the mixing-proportion and mean ability parameters under all simulated conditions, no matter whether there were two or three latent classes. For the two-class scenarios, the *RMSE*s for estimating the mixing-proportion parameters tended to decrease with the increase of either sample size or test length. However, this pattern was not observed with the three-class scenarios or with the recovery of the mean abilities. Given that both two- and three-class conditions considered the same sample size per class ($n = 250$ or 500) and test length ($J = 20$ or 30) conditions, parameter recovery results can also be compared across the $G = 2$ versus $G = 3$ scenarios. Hence, a comparison of Tables 1 and 2 reveals that the *RMSE*s for estimating the mixing-proportion parameters tended to decrease with the increase in the number of latent classes from two to three classes, except for one scenario (i.e., $N = 1000$, $J = 30$). This is, however, not the case with

Table 1 Bias and RMSE for recovering mixing-proportion and mean ability parameters when $G = 2$

N	J	Parameter	*Bias*	*RMSE*	Parameter	*Bias*	*RMSE*
500	20	π_1	−0.004	0.019	μ_1	−0.016	0.205
		π_2	0.004	0.019	μ_2	−0.085	0.196
	30	π_1	−0.003	0.013	μ_1	−0.003	0.013
		π_2	0.003	0.013	μ_2	0.003	0.013
1000	20	π_1	0.005	0.012	μ_1	0.116	0.197
		π_2	−0.005	0.012	μ_2	−0.039	0.146
	30	π_1	−0.001	0.011	μ_1	0.111	0.152
		π_2	0.001	0.011	μ_2	−0.089	0.150

Table 2 Bias and RMSE for recovering mixing-proportion and mean ability parameters when $G = 3$

N	J	Parameter	*Bias*	*RMSE*	Parameter	*Bias*	*RMSE*
750	20	π_1	−0.002	0.012	μ_1	0.074	0.242
		π_2	0.001	0.012	μ_2	−0.008	0.102
		π_3	0.001	0.010	μ_3	−0.026	0.260
	30	π_1	−0.002	0.008	μ_1	−0.292	0.333
		π_2	−0.002	0.010	μ_2	−0.034	0.133
		π_3	0.006	0.010	μ_3	0.190	0.293
1500	20	π_1	−0.001	0.010	μ_1	0.032	0.260
		π_2	−0.001	0.008	μ_2	−0.031	0.085
		π_3	0.002	0.007	μ_3	0.075	0.189
	30	π_1	−0.005	0.007	μ_1	−0.282	0.355
		π_2	0.010	0.012	μ_2	−0.025	0.062
		π_3	−0.005	0.008	μ_3	0.284	0.363

the mean ability parameters, whose *RMSE*s tended to increase when $G = 2$ increased to $G = 3$.

It is further noted that for the three-class scenarios, the accuracy of estimating the mean ability of the second latent class was better than that of the first or third latent class (see Table 2). In addition, the precision of the mean ability estimates for the second latent class improved with the increase in the sample size.

3.2 Item Parameter Recovery

The results for recovering the difficulty and discrimination parameters are summarized in Tables 3 and 4 for the two- and three-class conditions, respectively. These

Table 3 Average *Bias* and *RMSE* for recovering item parameters when $G = 2$

N	*J*	Parameter	*Bias*	*RMSE*	Parameter	*Bias*	*RMSE*
500	20	a_1	−0.074	0.397	b_1	0.396	0.626
		a_2	−0.063	0.400	b_2	−0.457	0.669
	30	a_1	−0.061	0.356	b_1	0.419	0.601
		a_2	−0.055	0.359	b_2	−0.493	0.691
1000	20	a_1	−0.014	0.298	b_1	0.447	0.638
		a_2	−0.076	0.339	b_2	−0.397	0.594
	30	a_1	−0.020	0.288	b_1	0.436	0.616
		a_2	−0.037	0.300	b_2	−0.382	0.609

Table 4 Average *Bias* and *RMSE* for recovering item parameters when $G = 3$

N	*J*	Parameter	*Bias*	*RMSE*	Parameter	*Bias*	*RMSE*
750	20	a_1	−0.054	0.409	b_1	0.386	0.522
		a_2	−0.049	0.469	b_2	0.057	0.421
		a_3	−0.053	0.443	b_3	−0.398	0.590
	30	a_1	−0.108	0.413	b_1	0.341	0.509
		a_2	−0.078	0.482	b_2	0.017	0.396
		a_3	−0.085	0.452	b_3	−0.375	0.545
1500	20	a_1	0.023	0.339	b_1	0.352	0.517
		a_2	−0.058	0.482	b_2	0.054	0.421
		a_3	−0.096	0.419	b_3	−0.311	0.499
	30	a_1	−0.058	0.383	b_1	0.377	0.558
		a_2	−0.071	0.420	b_2	0.035	0.379
		a_3	−0.088	0.356	b_3	−0.421	0.579

results indicate that with smaller average *bias* or *RMSE*, NUTS was more accurate in recovering the discrimination parameter than the difficulty parameter of the Mix2PL model for both classes in the two-class condition and for the first and third classes in the three-class condition.

The small negative values of the average *bias* for estimating the discrimination parameters suggest that they were slightly underestimated across all the simulated conditions except for one condition (i.e., $N = 1500$ and $J = 20$) where the discrimination for the first class was overestimated (see Table 4). For the two-class condition, the recovery of the discrimination parameters improved with the increase in sample size or test length, however, this pattern was not observed in the three-class condition, which has mixed results.

The difficulty parameters were consistently underestimated for the last latent class while overestimated for the other classes, no matter whether there were two or three classes. Also for the three-class condition, the recovery of the difficulty parameters

Table 5 Percent of correct classifications of individual persons

$G = 2$					$G = 3$				
N	J	Average	Min	Max	N	Average	J	Min	Max
500	20	90.96	74.40	97.20	750	69.65	20	65.20	81.60
	30	92.38	80.80	97.20		69.91	30	66.53	87.60
1000	20	93.55	82.80	96.10	1500	71.59	20	66.60	83.40
	30	94.44	86.50	97.20		75.13	30	64.20	90.73

in the second class, as indicated by the average values of *bias* and *RMSE*, was better than the recovery of those in the first or third class across the four data sizes.

In addition, a comparison of Tables 3 and 4 suggests that the average *RMSE*s for estimating the discrimination parameter tended to increase with the increase in the number of latent classes. On the hand, the *RMSE*s for estimating the difficulty parameters tended to decrease with the increase in the number of latent classes.

3.3 Class Membership Recovery

For the class membership, the percentages of correct classifications of individual persons were computed and displayed in Table 5, which suggests that NUTS was fairly accurate when the population consisted of two latent subpopulations. The average percentages of correct classifications, across the ten replications, for the four data sizes were 90.96, 92.38, 93.55, and 94.44. However, in the conditions where the population consisted of three latent subpopulations, the recovery was less accurate, where the average percentages of correct classifications for the four data sizes were 69.65, 69.91, 71.59, and 75.13. Moreover, the recovery of class memberships is apparently affected by sample size and test length. Specifically, the average percentage of correct classifications increased with an increase in sample size or test length, for both the two- and the three-class conditions.

4 Discussion and Conclusion

With Monte Carlo simulations, results of this study suggest that overall, NUTS performs well in recovering parameters for the Mix2PL model, including the class parameters (π_g and μ_g), item parameters (a_{jg} and b_{jg}), and class membership (g), although the recovery of the class membership of individual persons is not satisfactory for the three-class condition.

With respect to the effects of sample size or test length, they play a role in recovering the class membership no matter whether the generated data sets consisted of two or three latent subpopulations. This is consistent with previous research (e.g., Cho et al., 2013) where the proportion of correct classification of class membership increased with either sample size or test length. However, their effects on estimating other parameters in the Mix2PL model is not clear, as some patterns of recovery improvement with the increment of sample size and/or test length in the two-class condition are not observed in the three-class condition. For example, for the two-class condition, the accuracy of estimating the mixing-proportion parameters increases with the increase of either sample size or test length but this pattern is not observed with the three-class condition. This is possibly due to the increased complexity of the mixture item response theory (MixIRT) model with the increased number of latent classes. Adding one subpopulation may seem trivial, but it would result in a substantial increase in the number of parameters to be estimated. This complexity is further reflected in the estimation of person mean ability or item discrimination parameters, whose accuracy decreases with the increased number of classes. On the other hand, the recovery of the mixing proportions or individual item difficulties in the model is not seemingly affected by such added complexity. As a matter of fact, their *RMSE* values decrease when adding one more subpopulation. This reduction is due to the fact that the magnitude of *RMSE* depends on the unit/scale of the parameter. For instance, the mixing proportion is larger for the two-class condition ($\pi_g = 0.5$) than the three-class condition ($\pi_g = 0.33$), and hence the *RMSEs* tend to be larger with the two-class condition. This is certainly a limitation of using *RMSE* for evaluating the accuracy in recovering model parameters in this study. Future studies shall consider other measures, such as the relative *RMSE* or normalized *RMSE* that are free from the scale of the parameters.

The finding that the discrimination parameter is better recovered than the difficulty parameter in the MixIRT model (based on the comparison of average *RMSE*/*bias* values) agrees with Chang (2017), who focused on the estimation of the conventional IRT model using NUTS and Gibbs sampling. However, it does not agree with findings from studies on fitting some other IRT models with non-Bayesian estimations (e.g., Batley & Boss, 1993; Kang & Cohen, 2007) although the same *RMSE* criterion has been used. Given the limitation of *RMSE* as noted previously, further studies are needed to direct the trend of such comparisons. In addition, results based on the three-class situation suggest that the item difficulty or the class mean ability parameters are estimated more accurately for the second class than for the first or third class. This is likely due to the choice of the simulated person ability and item difficulty parameters for each of the three latent classes. Specifically, the generated person abilities for the second class (i.e., $\theta_2 \sim N(0, 1)$) coincides with the generated item difficulty (i.e., $b_2 \sim U(-0.5, 0.5)$) for that class. However, the generated person abilities for the first class (i.e., $\theta_1 \sim N(-4, 1)$) is quite distant from the generated item difficulty (i.e., $b_2 \sim U(-2, -0.5)$) for that class, such that the average person ability (i.e., -4) is 2.75 standard deviations lower than the average item difficulty (i.e., -1.25). Similarly, the generated person ability for the third class (i.e., $\theta_3 \sim N(4, 1)$) is quit distant from the generated item difficulty (i.e., $b_2 \sim U(0.5, 2)$) for that class, such that the average person ability (i.e., 4) is 2.75 standard deviations higher than the average item difficulty (i.e., 1.5). Thus, in order to obtain more accurate estimates of the person mean ability and item difficulty parameters for the first class, more easy items should be added. On the other hand, in order to obtain more precise estimations of the person mean ability and item difficulty parameters for the third class, more difficult items should be added.

This study provides empirical evidence on the performance of NUTS in fitting MixIRT models. It also shows that researchers and practitioners in educational and psychological measurement can use NUTS in estimating parameters of complex IRT models such as MixIRT models. However, conclusions that are made in the present study are based on the simulated conditions and cannot be generalized to other conditions. Therefore, for future studies, additional test conditions need to be explored such as unequal mixing proportions, small sample size, and short test length. Given the computational expense of fitting NUTS to the complex Mix2PL model, this study only used 10 replications for each experimental condition. However, as suggested by Harwell, Stone, Hsu, & Kirisci, (1996), a minimum of 25 replications is recommended for typical Monte Carlo studies in IRT modeling. Additional studies with similar experimental conditions are needed before one can conclude about the use of the algorithm with fitting the Mix2PL model and further the effects of sample size, test length, and number of classes on estimating the model. In addition, this study focused on the dichotomous Mix2PL model. Future studies may consider evaluating the performance of NUTS using other dichotomous MixIRT models such as the Mix1PL model or the Mix3PL models, or using MixIRT models with polytomous categories such as a mixture version of Bock's (1972) nominal response model or a mixture version of Masters's (1982) partial credit model. Moreover, this study considered certain population distributions and difficulty ranges. Additional studies are necessary to consider other person distributions and/or other ranges for item difficulty parameters to decide on the test condition that leads to more accurate estimates for all classes. Future studies are also needed to decide on the optimal number of persons and/or items for more accurate estimations of class membership in conditions where the population includes three or more subpopulations for any given class size. Finally, findings from this study are based on simulated conditions where the true parameters are known, Future studies may adopt NUTS algorithms to fit the Mix2PL models to real data and examine how NUTS performs in real test situations.

References

Batley, R.-M., & Boss, M. W. (1993). The effects on parameter estimation of correlated dimensions and a distribution-restricted trait in a multidimensional item response model. *Applied Psychological Measurement, 17*(2), 131–141. https://doi.org/10.1177/014662169301700203.

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology, 6*(2), 258–276.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29–51.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*(4), 331–348.

Chang, M. (2017). *A comparison of two MCMC algorithms for estimating the 2PL IRT models*. Doctoral: Southern Illinois University.

Cho, S., Cohen, A., & Kim, S. (2013). Markov chain Monte Carlo estimation of a mixture item response theory model. *Journal of Statistical Computation and Simulation, 83*(2), 278–306.

Choi, Y., Alexeev, N., & Cohen, A. S. (2015). Differential item functioning analysis using a mixture 3-parameter logistic model with a covariate on the TIMSS 2007 mathematics test. *International Journal of Testing, 15*(3), 239–253. https://doi.org/10.1080/15305058.2015.1007241.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement Summer, 42*(2), 133–148.

De Ayala, R. J., Kim, S. H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: a mixture distribution conceptualization. *International Journal of Testing, 2*(3&4), 243–276.

de la Torre, J., Stark, S., & Chernyshenko, O. S. (2006). Markov chain Monte Carlo estimation of item parameters for the generalized graded unfolding model. *Applied Psychological Measurement, 30*(3), 216–232. https://doi.org/10.1177/0146621605282772.

Duane, S., Kennedy, A., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B, 195,* 216–222. https://doi.org/10.1016/0370-2693(87)91197-X.

Finch, W. H., & French, B. F. (2012). Parameter estimation with mixture item response theory models: A Monte Carlo comparison of maximum likelihood and Bayesian methods. *Journal of Modern Applied Statistical Methods, 11*(1), 167–178.

Gelfand, A. E., & Sahu, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *JASA, 94*(445), 247–253. https://doi.org/10.2307/2669699.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Florida: CRC Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat Sci, 7*(4), 457–472.

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

Grant, R. L., Furr, D. C., Carpenter, B., & Gelman, A. (2016). Fitting Bayesian item response models in Stata and Stan. *The Stata Journal, 17*(2), 343–357. https://arxiv.org/abs/1601.03443. Accessed 18 Apr 2018.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101–125. https://doi.org/10.1177/014662169602000201.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57*(1), 97–109. https://doi.org/10.1093/biomet/57.1.97.

Hoffman, M. D., & Gelman, A. (2011). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15*(2), 1593–1624.

Huang, H. (2016). Mixture random-effect IRT models for controlling extreme response style on rating scales. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.01706.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31*(4), 331–358. https://doi.org/10.1177/0146621606292213.

Kim, S.-H. (2007). Some posterior standard deviations in item response theory. *Educational and Psychological Measurement, 67*(2), 258–279. https://doi.org/10.1177/00131644070670020501.

Li, F., Cohen, A., Kim, S., & Cho, S. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement, 33*(5), 353–373. https://doi.org/10.1177/0146621608326422.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (2nd ed.). New Jersey: Hillsdale.

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

Luo, Y., & Jiao, H. (2017). Using the Stan program for Bayesian item response theory. *Educational and Psychological Measurement*, 1–25. https://doi.org/10.1177/0013164417693666.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2010). Improvement in detection of differential item functioning using a mixture item response theory model. *Multivariate Behavioral Research, 45*(6), 975–999. https://doi.org/10.1080/00273171.2010.533047.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association, 44*(247), 335–341.

Meyer, J. P. (2010). A mixture Rasch model with Item response time components. *Applied Psychological Measurement, 34*(7), 521–538. https://doi.org/10.1177/0146621609355451.

Mroch, A. A., Bolt, D. M., & Wollack, J. A. (2005). *A new multi-class mixture Rasch model for test speededness*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Quebe, April 2005.

Neal, R. M. (1992). *An improved acceptance procedure for the hybrid Monte Carlo algorithm*. Retrieved from arXiv preprint https://arxiv.org/abs/hep-lat/9208011.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, & X. Meng (Eds.), *Handbook of Markov chain Monte Carlo* (pp. 113–162). Florida: CRC Press.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(1), 1–18.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (2nd ed.). Danmark: Danmarks Paedagogiske Institute.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*(3), 271–282. https://doi.org/10.1177/014662169001400305.

Samuelsen, K. (2005). *Examining differential item functioning from a latent class perspective* (Dissertation). University of Maryland.

Shea, C. A. (2013). *Using a mixture IRT model to understand English learner performance on large-scale assessments* (Dissertation). University of Massachusetts.

Stan Development Team. (2017). *Stan modeling language users guide and reference manual, version 2.17.0*. http://mc-stan.org. Accessed 8 Feb 2018.

van der Linden, Wd, & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer.

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y. S. (2002). Recovery of item parameters in the nominal response model: a comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 26*(3), 339–352. https://doi.org/10.1177/0146621602026003007.

Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement, 40,* 307–330.

Wu, X., Sawatzky, R., Hopman, W., Mayo, N., Sajobi, T. T., Liu, J., … Lix, L. M. (2017). *Latent variable mixture models to test for differential item functioning: a population-based analysis. Health and Quality of Life Outcomes, 15*. https://doi.org/10.1186/s12955-017-0674-0.

Zhu, L., Robinson, S. E., & Torenvlied, R. (2015). A Bayesian approach to measurement bias in networking studies. *The American Review of Public Administration, 45*(5), 542–564. https://doi.org/10.1177/0275074014524299.

Controlling Acquiescence Bias with Multidimensional IRT Modeling

Ricardo Primi, Nelson Hauck-Filho, Felipe Valentini, Daniel Santos and Carl F. Falk

Abstract Acquiescence is a commonly observed response style that may distort respondent scores. One approach to control for acquiescence involves creating a balanced scale and computing sum scores. Other model-based approaches may explicitly include an acquiescence factor as part of a factor analysis or multidimensional item response model. Under certain assumptions, both approaches may result in acquiescence-controlled scores for each respondent. However, the validity of the resulting scores is one issue that is sometimes ignored. In this paper, we present an application of these approaches under both balanced and unbalanced scales, and we report changes in criterion validity and respondent scores.

Keywords Acquiescence bias · Item response modeling

This article is part of a research project supported by the Ayrton Senna Foundation. The first author also received a scholarship from the National Council on Scientific and Technological Development (CNPq). Correspondence should be addressed to: Ricardo Primi, Universidade São Francisco, Rua Waldemar César da Silveira, 105, Campinas, São Paulo, CEP 13045-510; e-mail: rprimi@mac.com.

R. Primi (✉) · D. Santos
Ayrton Senna Institute, Universidade São Francisco, and EduLab21, Campinas, Brazil
e-mail: rprimi@mac.com

D. Santos
e-mail: daniel.ddsantos@gmail.com

N. Hauck-Filho · F. Valentini
Universidade São Francisco, Campinas, Brazil
e-mail: hauck.nf@gmail.com

F. Valentini
e-mail: valentini.felipe@gmail.com

C. F. Falk
McGill University, Montreal, Canada
e-mail: carl.falk@MCGILL.CA

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_4

1 Introduction

1.1 Large-Scale Assessment of Socioemotional Skills and the Self-report Method

Evidence has consistently indicated that socioemotional skills can predict many life outcomes (Ozer & Benet-Martínez, 2006), including job-related variables (Heckman, Stixrud, & Urzua, 2006), quality of life (Huang et al., 2017), psychopathology (Samuel & Widiger, 2008), and physical health (Allen, Walter, & McDermott, 2017). Among students, such skills have been associated with academic performance even when partialling out intelligence (Poropat, 2014), perhaps because these skills foster multiple learning strategies and positive self-beliefs (Zhang & Ziegler, 2018). Considering that such individual differences not only change over time (Soto, John, Gosling, & Potter, 2011), but can also be enhanced via school-based interventions (Lipnevich, Preckel, & Roberts, 2016), they represent key variables to modern national education policies.

Although many strategies exist for the assessment of socioemotional skills among students, the self-report method is recommended because it is simple, easy, and has a low cost compared to alternative techniques (Kyllonen, Lipnevich, Burrus, & Roberts, 2014). One recently published self-report inventory designed for the assessment of non-cognitive skills among students is SENNA (Primi, Santos, John, & De Fruyt, 2016). It contains 18 self-report scales using 5-point Likert-type items and provides researchers and public agencies with information on five broad dimensions of socioemotional skills: Open-mindedness (O), Conscientious Self-Management (C), Engaging with others (E), Amity (A), and Negative-Emotion Regulation (N) (John, Naumann, & Soto, 2008).

1.2 Self-report Method and Response Styles

Although the self-report method has many merits, it does not result in error-free information about respondents. Scores calculated on self-report data may be contaminated by random error or by systematic components unrelated to the trait of interest. Systematic biases include “response styles” (Paulhus, 1991) or “method variance” (McCrae, 2018). Response styles (RS) represent individual differences in the usage of response scales. For instance, some respondents will tend to manifest their agreement or disagreement with the content of an item by choosing the extremes of the Likert scale, while others will systematically avoid extremes. RS represent relatively stable individual differences (Weijters, Geuens, & Schillewaert, 2010; Wetzel, Lüdtke, Zettler, & Böhnke, 2015) and may account for up to 40% of item variance (McCrae, 2018). When separating trait and state components in repeated measures designs, response styles seem to be responsible for up to 59% of systematic state variance (Wetzel et al., 2015). By adding nuisance variance to the data, RS can impair the validity and reliability of test scores (Ziegler, 2015).

1.3 Acquiescence and the Assessment of Socioemotional Skills

Acquiescence (ACQ) is one response style that deserves closer attention in the self-report assessment of socioemotional skills among youths. ACQ refers to a tendency to agree with items at the expense of their content (Paulhus, 1991). For instance, a student might indicate that he or she agrees (e.g., "4" on a 5-point Likert scale ranging from 1 = strongly disagree to 5 = strongly agree) with two items such as "I am an introvert" and "I am an extravert." Of course, such a response pattern is semantically contradictory, and it indicates agreement in detriment to consistency. In some cases, ACQ may reflect cognitive simplicity (Knowles & Nathan, 1997) as it occurs more often among under-educated people (Meisenberg & Williams, 2008), older adults (Weijters et al., 2010) and younger children and adolescents (Soto, John, Gosling, & Potter, 2008).

With respect to self-reports of socioemotional skills, ACQ has the potential to diminish correlations between semantically opposite items, creating method factors among negatively worded items (Kam & Meyer, 2015). ACQ can also increase correlations among items capturing unrelated traits (Soto et al., 2008). Accordingly, factor structure distortions are very likely to occur in the presence of ACQ. In a simulation study, ACQ caused classical parallel analysis and Hull methods to overestimate the number of factors to retain, and MAP and permutation parallel analysis to underestimate it (Valentini, 2017).

Moreover, ACQ can attenuate external validity (Mirowsky & Ross, 1991). ACQ tends to inflate scores of scales composed of mostly positively worded items. Thus, ACQ might impact the validity of a scale in a manner proportional to the amount of positively- and negatively-keyed items. At the same time, ACQ is often negatively related to achievement, suggesting that high ACQ can be explained in part by low language skills. Therefore, the criterion validity of socio-emotional skills may be suppressed by ACQ. In real data and using a classical scoring approach, Primi, De Fruyt, Santos, Antonoplis, and John (2018) found that partialling out ACQ resulted in disattenuated associations of socioemotional skills with achievement tests of language (from .13 to .21) and math (from .11 to .17).

1.4 Controlling for Acquiescence

One traditional way of controlling for ACQ is to create a balanced scale in which each positively worded item is paired with an antonym (a negatively worded item), such as: *I am often talkative /I am often quiet*. On balanced scales, it is expected that subjects will give mirrored responses to antonym pairs (e.g., 5-1, 4-2, 3-3, 2-4 and 1-5 on a 5-point Likert-type item). If the response pattern of subject *j* is semantically consistent, then the average of subject *j*'s item responses before reverse coding will be the midpoint of the response options (in this case, 3; Soto et al., 2008). The person's average of the item responses before reverse coding negative items is the classical index of ACQ (acq_j).

Under certain assumptions (e.g., positively and negatively worded items are on average equally vulnerable to ACQ; Savalei & Falk, 2014a, b), classical scoring procedures will result in unbiased estimates of the respondents' scores. In essence, the effect of ACQ on positive and negatively worded items "cancels out" when computing a total score. For example, Primi et al. (2019) shows that scr_j, the classical average score of subject *j* on a balanced scale (with a 5-point Likert type item scored from 1 to 5), can be written as:

$$scr_j = 3 + \frac{1}{2}\left(\frac{\sum_{i=1}^{k^{(p)}} x_{ij}^{(p)}}{k^{(p)}} - \frac{\sum_{i=1}^{k^{(n)}} x_{ij}^{(n)}}{k^{(n)}}\right)$$

where $k^{(p)}$ equals the number of positive items, $k^{(n)}$ is the number of negative items, $x_{ij}^{(p)}$ and $x_{ij}^{(n)}$ are subject j's original responses (before reverse coding) on positive item i, and negative item i, respectively. Inside parentheses, the classical score is a function of the difference between the average agreement with positive versus negative items. The more inconsistent the responses to antonym items are, the more the term in parentheses will tend towards zero. Semantically consistent responses, however, will tend to result in either larger or smaller scr_j, depending on the subject's standing on the trait.

In unbalanced scales (i.e., $k^{(p)} \neq k^{(n)}$), classical scores may not be fully corrected and ACQ will not fully cancel out. In such a case, a form of within-person centering (or ipsatization) is sometimes recommended to control for ACQ (e.g., Soto et al., 2008). In the first step, an ACQ index (acq_j) is calculated as the average of only antonym pairs of items. Next, ACQ is removed from the raw item scores $(x_{ij} - acq_j)$. Raw scores for the reverse-keyed items are then multiplied by -1, and scale scores are obtained by averaging these items with those of the positively worded items.

1.5 Item Response Theory with Questionnaires and Acquiescence

Item response theory (IRT) models are routinely used when scaling constructs derived from questionnaires in large-scale educational assessments. While much is known about the effect of ACQ in balanced and unbalanced classical scores (Ten Berge, 1999; Primi et al., 2019), less is understood about the effect of ACQ on latent trait scores estimated via IRT. Since it is known that ACQ, even with a balanced scale, may contaminate the covariance structure when performing linear factor analysis (e.g., Savalei & Falk, 2014a, b), we conjecture that IRT-based models may also be vulnerable to the effects of ACQ. For instance, the graded response model and generalized partial credit models are commonly used IRT models for the analysis of ordered polytomous responses (De Ayala, 2009), but may not automatically correct for ACQ. There are, however, a number of model-based approaches that could be used to control for ACQ, such as those based on the random intercept model (Billiet & McClendon, 2000; Cai, 2010; Maydeu-Olivares & Coffman, 2006; Maydeu-Olivares, & Steenkamp, 2018).

Although we provide some details on these models later, some key questions emerge regarding the consequences of ACQ regardless of the method used. Simulations and analytical proofs are useful for studying whether a modeling approach can recover population parameters or results in bias, as well as the consequences of fitting a misspecified model. In practice, however, we never know the true model and whether a more complex modeling approach fits the data better because it is a better approximation to reality or because it is fitting noise. Supposing that we are interested in using self-management scores to predict an objective real-world outcome, we may wonder about the consequences of ignoring ACQ or using a specialized approach to control for it. For example, how does the use of one model versus another affect the validity of IRT scores? Are there differences if questionnaires are balanced or unbalanced? Are there any differences in scoring bias when comparing classical and IRT-based approaches? We therefore present an empirical study comparing the criterion validity of classical scores against four IRT approaches.

2 Method

Our main goal was to explore the criterion validity of self-management scores estimated via IRT. Previous research with classical scores suggests that ACQ suppresses criterion validity, and that ACQ-controlled scores show relatively higher validity (Primi, Santos, De Fruyt, & John, 2018). In the present study, we calculated scores via IRT, and then explored their criterion validity. We wanted to examine if classical scores are similar to ACQ-controlled trait estimates. We also compared these findings on a balanced versus an unbalanced item set.

2.1 Data

We reanalyzed data from Primi et al. (2018). Data comprised of 12,987 adolescents (52.7% female) from grades 7, 9, and 10, who ranged in age from 12 to 20 years (*M* = 16, *SD* = 1.85). Participants were regular students attending 425 public schools located in 216 cities of the state of Sao Paulo. Students completed SENNA as part of a reading literacy program developed by the Ayrton Senna Institute and in partnership with the state secretariat.

2.2 Instruments

We focused on the 45-item Conscientious Self-Management Scale (C) from the SENNA inventory (Primi et al., 2018). The scale contains 30 antonym pairs, 15 positively-keyed and 15 negatively-keyed items, with an additional 15 positively-keyed items. The scale is therefore unbalanced. In what follows, we performed the analyses twice: Once on the 30 antonym pairs (the balanced item set), and a second time on the complete 45-item set (the unbalanced item set). Students responded using a 5-point scale. We also had two measures of students' academic achievement: standardized assessments for language and math (SARESP—Assessment of Educational Achievement at São Paulo State, in Portuguese—see http://saresp.fde.sp.gov.br). These scores were used as criterion measures.

2.3 Data Analysis and Multidimensional IRT Modeling

In synthesis, the study design crossed two features: (a) two types of item sets: Balanced versus unbalanced; and (b) five psychometric models to calculate scores: Classical, unidimensional IRT via a graded response model (GRM), a unidimensional partial credit model (PCM; e.g., see De Ayala, 2009; Embretson, & Reise, 2000), and two multidimensional IRT models that were an adaptation of the random intercept model but based on either the GRM or PCM. Our main focus was the correlation between self-management and standardized achievement in language and math.

When calculating classical scores, we obtained original scores (*Raw ave*) that are simply the average of item responses after reverse coding negative items (equivalent to computation of scr_j). We also calculated classical ACQ-controlled scores (*ACQ cntr*) using the procedure advocated by Soto et al. (2008) for unbalanced items as mentioned earlier in our manuscript, along with an acquiescence index (*ACQ*) via average endorsement of the 15 antonym pairs before reverse coding. Note that in the case of a balanced scale, *Raw ave* and *ACQ cntr* are equivalent; these indices differ only for unbalanced scales.

To understand the two random intercept models, consider boundary discrimination functions for the GRM as follows

$$P_{ri} = \frac{1}{1 + \exp\bigl(-\bigl(a_{1i}\theta_j + a_{2i}\zeta_j + c_{ri}\bigr)\bigr)}$$

where P_{ri} is short-hand for the probability of endorsing category r or higher for item i. For each 5-point Likert-type item there will be four of these equations modeling the transitions 1 versus 2345, 12 versus 345, 123 versus 34, and 1234 versus 5. a_{1i} is the discrimination for item i on the substantive trait, θ_j, and a_{2i} is a set of fixed weights for item i associated with item wording and designed to capture ACQ. Values of a_{2i} were fixed to 1 if the item was positively worded, and -1 if the item was negatively worded. With this fixed set of weights, ζ_j represents ACQ. Finally, c_{ir} represents an intercept term. This model is similar to what Maydeu-Olivares and Steenkamp (2018) named the compensatory random-intercept model (see also Cai, 2010).

To estimate the model, we freed item discriminations (a_{1i}), and constrained the trait variance to 1. Since we fixed all a_{2i} parameters, we freed the variance of the acquiescence factor, ζ_j, and fixed the covariance between trait and acquiescence to zero for identification. We also estimated a second model with all specifications similar to the GRM but using a multidimensional PCM. This model fixed item discriminations to 1, and estimated substantive trait variance. After calibrating item parameters, we estimated subject factor scores using the *Expected* a Posteriori (EAP) algorithm. Trait and acquiescence scores were named *GRM f1* and *GRM f2*, respectively, for the GRM and *PCM f1* and *PCM f2*, respectively, for the PCM. (Chalmers, 2012)

3 Results

Table 1 shows descriptive statistics and criterion validity of the distinct types of scores investigated here. Whereas the upper half of the table shows scores calculated with a set of items balanced with respect to item wording, the lower half displays the same set of scores but calculated using the unbalanced set of items. The last two columns show zero-order correlations of various scores with standardized achievement in language and math.

Some key points are worth noticing in Table 1. First, considering classical scores in the balanced condition, we found that self-management was positively associated with achievement in magnitudes consistent with previous literature (see Poropat, 2009), while acquiescence tended to be negatively associated with achievement (Mirowsky & Ross, 1991). Second, *Raw ave* and *ACQ cntr* had the same association with achievement ($r = .22$ and .18 for language and math, respectively). Considering the unbalanced item set, *Raw ave* showed smaller correlations with achievement ($r = .16$ and .14) than did acquiescence-controlled scores, *ACQ cntr* ($r = .20$ and .16). This result is likely a consequence of the suppression effect of ACQ discussed earlier (see Primi et al., 2018). The negative correlation of *ACQ* with standardized achievement in language ($r = -.12$) was slightly stronger than its correlation with math ($r = -.08$), corroborating the idea that ACQ is associated with poor language skills.

When we consider IRT estimated scores from the balanced set of items, we also found a positive correlation between trait and achievement, but a negative correlation between acquiescence and achievement. It is interesting to note that only the PCM had validity coefficients that were of a similar magnitude as classical scores. On the one hand, this is not surprising as sum scores are a sufficient statistic for estimating

Table 1 Descriptive statistics and criterion validity of various scores based on classical, partial credit, graded response model and random intercept multidimensional IRT models

Variables	M	SD	Min	Max	Correlation	
					Lang.	Math
Balanced scale						
Classical scores						
Raw ave		0.57 3.55	1.13	5.00	0.22	0.18
ACQ cntr		0.57 0.55	−1.87	2.00	0.22	0.18
ACQ		0.35 2.95	1.00	5.00	−0.12	−0.08
Unidimensional IRT						
GRM[a]		0.96 0.00	−4.40	3.28	0.17	0.14
PCM[b]		0.54 0.00	−2.39	2.05	0.22	0.18
Random intercept MIRT						
GRM f1[c]		0.96 −0.01	−4.40	3.29	0.18	0.15
GRM f2		0.59 0.00	−3.58	3.67	−0.11	−0.07
PCM f1[4]		0.60 0.00	−2.67	2.28	0.21	0.17
PCM f2		0.30 0.00	−1.80	1.78	−0.10	- 0.07
Unbalanced scale						
Classical scores						
Raw ave		0.58 3.53	1.07	5.00	0.16	0.14
ACQ cntr		0.58 0.57	−2.00	2.11	0.20	0.16
Unidimensional IRT						
GRM[e]		0.98 0.00	−4.80	3.58	0.12	0.10
PCM[f]		0.61 0.01	−2.93	2.57	0.19	0.15
Random intercept MIRT						
GRM f1[g]		0.97 −0.01	−4.72	3.59	0.16	0.13
GRM f2		0.63 0.00	−3.99	4.08	−0.12	−0.08
PCM f1[h]		0.66 0.00	−3.09	2.71	0.20	0.16
PCM f2		0.35 0.00	−2.23	2.19	−0.10	−0.07

Note: *Raw ave* classical scores calculated via average item endorsing after reversing negatively phrased items; *ACQ cntr* classical scores controlled for acquiescence using the procedure by Soto et al. (2008); *ACQ* classical acquiescence index calculated via average endorsement of 15 antonym pairs of items before reversing negatively phrased items; *PCM* IRT estimated scores based on the partial credit model; *GRM* IRT estimated scores based on the graded response model; *GRM f1* and *GRM f2* trait and acquiescence scores estimated from the random intercept graded response model; *PCM f1* and *PCM f2* trait and acquiescence scores estimated from the random intercept partial credit model. Fit indices were: [a]*CFI* = .68, *RMSEA* = .08, *AIC* = 1,062,104, *BIC* = 1,063,008; [b]*CFI* = .73, *RMSEA* = .07, *AIC* = 1,043,494, *BIC* = 1,044,615; [c]*CFI* = .82, *RMSEA* = .06, *AIC* = 1,022,051, *BIC* = 1,023,179; [d]*CFI* = .77, *RMSEA* = .07, *AIC* = 1,042,179, *BIC* = 1,043,091; [e]*CFI* = .73, *RMSEA* = .08, *AIC* = 1,513,148, *BIC* = 1,514,829; [f]*CFI* = .78, *RMSEA* = .07, *AIC* = 1,554,601, *BIC* = 1,555,954; [g]*CFI* = .84, *RMSEA* = .06, *AIC* = 1,485,955, *BIC* = 1,487,643; [h]*CFI* = .82, *RMSEA* = .06, *AIC* = 1,5182,58, *BIC* = 1,519,618 the PCM (De Ayala, 2009), but surprising given that the GRM is often described as a more realistic model for data. Validity coefficients from the other three models (GRM and both random intercept models) were similar in magnitude, but were slightly lower than the validity of classical scores.

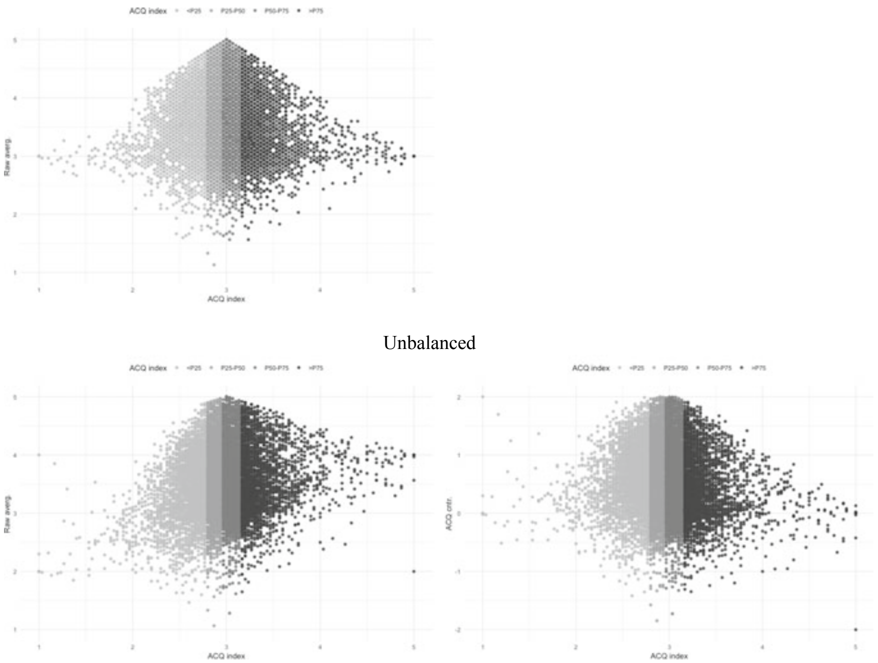
Although balanced scales have an equal number of items for each pole (15 items), we still found a difference in item discrimination under GRM across positively (1.27 on average) versus negatively worded (1.03 on average) items, which might yield an imbalance in the contribution of these items to EAP scores. Since the positive trait pole was favored, the correction process also becomes unbalanced, and is no longer similar to what occurred with classical scores. For instance, *Raw ave* correlated negatively with *ACQ* ($r = -.05$) in the balanced condition while *GRM* correlated positively with *ACQ* and *GRM f2* (r =.09 and .07 respectively) This indicates that the estimation of IRT scores may be slightly biased by ACQ even in balanced scales due to differences in the discrimination between positively- and negatively-keyed items. For instance, when discrimination is constrained equal across items (i.e., for the random intercept PCM), then the correlation between ACQ and the estimated trait becomes $r = -.04$.

For score estimates from the unbalanced item set, we found some noticeable differences. Since there were more positively worded items and they had higher discrimination (1.52 on average) than negatively worded items (0.79 on average) under GRM, this might lead to an even stronger positive association of ACQ with trait scores. In fact, classical acquiescence scores (*ACQ*) were positively correlated with *Raw ave*, $r = .18$, *PCM*, $r = .10$, and *GRM*, $r = .28$. By contrast, correlations between *ACQ* and self-management were lower when compared to EAP scores from the random intercept models, *GRM f1*, $r = .03$, and *PCM f1*, $r = .01$.

Score inflation due to ACQ tended to suppress the correlation between self-management and achievement. We see that the uncontrolled score *Raw ave* ($r =$ .16 and .14 for language and math, respectively) had lower validity coefficients than *ACQ cntr* ($r = .20$ and .16). The random intercept GRM, *GRM f1*, had better coefficients ($r = .16$ and .13) than the *GRM* ($r = .12$ and .10). Rasch models tended to have better validities, as the unidimensional *PCM* ($r = .19$ and .15) and random intercept PCM, *PCM f1* ($r = .20$, .16), had the best validity coefficients of any IRT model. Overall, it is possible that this result indicates that random intercept models are producing scores that may be better controlling for ACQ.

Figures 1 and 2 show the effect of ACQ correction on scores. The upper part of Fig. 1 shows the relationship between *ACQ* (x-axis) and *Raw ave* (y-axis) for the balanced item set. When the ACQ index was near 3, scores had the full amplitude of variation from 1 to 5. As subjects tended to respond inconsistently, that is, tended to have *ACQ* > 3 or *ACQ* < 3, score variation was reduced. When agreeing was not completely consistent, scores were corrected towards the scale's center.

The lower part of Fig. 1 shows what happens in the unbalanced item set. We see the relationship between *ACQ* (x-axis) and *Raw ave* (y-axis) on the left, and between *ACQ* and *ACQ cntr* on the right (y-axis). In all graphs, we see a diamond shape characterizing the ACQ correction, with an important difference. Original scores (*Raw ave*) were correlated positively with *ACQ* ($r = .18$), but ACQ controlled

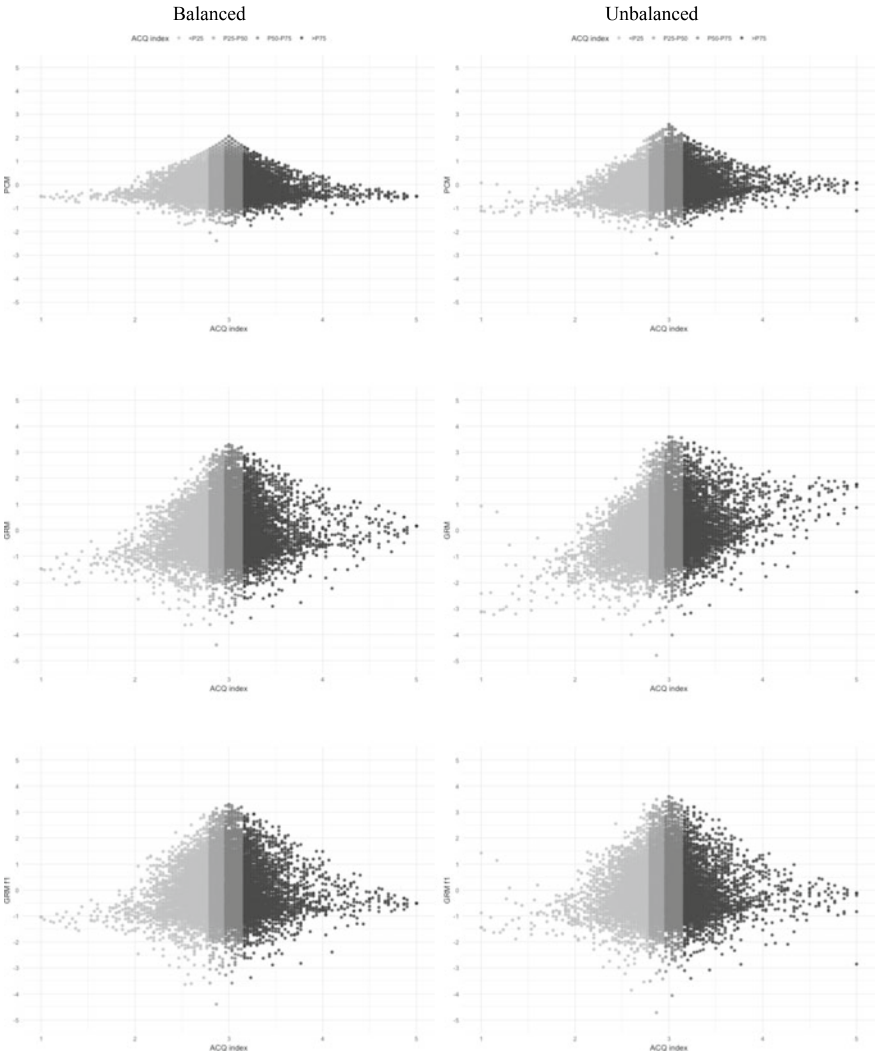


Unbalanced

Fig. 1 Effects of acquiescence correction in classical scores of balanced scales (upper panel) versus unbalanced scales (lower panel). Scores on *y*-axis and ACQ indexes in *x*-axis

scores (*ACQ cntr*) were slightly negatively correlated with *ACQ* ($r = -.08$). Because the scale had more positively than negatively worded items, the correction process tended to produce the opposite effect, lowering high scores and increasing low scores if subjects exhibited ACQ or disacquiescence, respectively. This impacts validity coefficients because, in theory, ACQ is partialled out of *ACQ cntr*.

Figure 2 shows what happens with IRT scores, with *ACQ* always on the x-axis. The left columns show plots for the balanced item set, and the right columns for the unbalanced item set. On the y-axis, the upper panels represent *PCM*, the middle panels *GRM* and the lower panels *GRM f1*. We see patterns similar to what is shown in Fig. 1. In that scores may be corrected for ACQ. Nevertheless, we observe some variability among methods in the amount that scores are confounded with ACQ. There is no confounding for the *PCM* under the balanced item set ($r = -.04$), but some confounding under the unbalanced item set ($r = .10$). The *GRM* was slightly confounded in the balanced item set ($r = .09$), but much more confounded in the unbalanced item set ($r = .28$). Finally, random intercept models were less confounded. For example, for both balanced and unbalanced item sets, *GRM f1* correlated with *ACQ* near zero, $r = .03$.


Balanced

Unbalanced

Fig. 2 Effects of acquiescence correction in IRT scores (partial credit—*PCM*, graded response—*GRM* and random intercept *GRM f1*) of balanced scales (left column) versus unbalanced scales (right column). Scores on *y*-axis and ACQ indexes in *x*-axis

4 Discussion

Acquiescence can negatively affect the criterion validity of self-report instruments. Balanced scales or acquiescence-controlled scores for unbalanced scales are ways to improve score validity (Mirowsky, & Ross, 1991; Primi et al., 2018; Soto & John, 2019). But less is known about ACQ corrections and the validity of IRT scores when scales are composed of both positively and negatively worded items. We tested five approaches spanning classical scoring, traditional IRT models and multidimensional IRT models based on the random intercept model (Billiet & McClendon, 2000; Cai, 2010; Maydeu-Olivares & Coffman, 2006; Maydeu-Olivares, & Steenkamp, 2018; Primi et al., 2018). The two modified versions of the random intercept models added an extra factor to explicitly model ACQ, and these were based on the GRM and PCM. These models produced ACQ-controlled IRT trait scores and also IRT ACQ index scores. The best of these models was the random intercept PCM.

We found that ignoring the possibility of ACQ is the worst-case scenario in terms of criterion validity. In balanced scales, the unidimensional PCM performed better than the GRM. With unbalanced scales, unidimensional GRM scores had the worst criterion validity. We suspect that either different item loadings for the GRM are picking up on some misspecification (lack of modeling ACQ) or that unique item content is important for criterion validity and is more equally considered under the PCM.

References

Allen, M. S., Walter, E. E., & McDermott, M. S. (2017). Personality and sedentary behavior: A systematic review and meta-analysis. *Health Psychology, 36*(3), 255–263. https://doi.org/10.1037/hea0000429.

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling, 7*(4), 608–628. https://doi.org/10.1207/S15328007SEM0704_5.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75,* 581–612. https://doi.org/10.1007/s11336-010-9178-0.

Chalmers, R. P. (2012). MIRT: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. Retrieved from http://www.jstatsoft.org/v48/i06/.

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

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics, 24*(3), 411–482. https://doi.org/10.1086/504455.

Huang, I.-C., Lee, J. L., Ketheeswaran, P., Jones, C. M., Revicki, D. A., & Wu, A. W. (2017). Does personality affect health-related quality of life? A systematic review. *PLOS ONE, 12*(3), e0173806. https://doi.org/10.1371/journal.pone.0173806.

John, O., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big-Five trait taxonomy: History, measurement, and conceptual issues. In O. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114–158). New York: Guilford Press.

Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality. *Organizational Research Methods, 18*(3), 512–541. https://doi.org/10.1177/1094428115571894.

Knowles, E. S., & Nathan, K. T. (1997). Acquiescent responding in self-reports: cognitive style or social concern? *Journal of Research in Personality, 31*(2), 293–301. https://doi.org/10.1006/jrpe.1997.2180.

Kyllonen, P. C., Lipnevich, A. A., Burrus, J., & Roberts, R. D. (2014). Personality, motivation, and college readiness: A prospectus for assessment and development. *ETS Research Report Series, 2014*(1), 1–48. https://doi.org/10.1002/ets2.12004.

Lipnevich, A. A., Preckel, F., & Roberts, R. D. (2016). *Psychosocial skills and school systems in the 21th century*. New York: Springer.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods, 11*(4), 344–362. https://doi.org/10.1037/1082-989X.11.4.344.

Maydeu-Olivares, A., & Steenkamp, J. E. M. (2018). *An integrated procedure to control for common method variance in survey data using random intercept factor analysis models*. https://www.academia.edu/36641946/An_integrated_procedure_to_control_for_common_method_variance_in_survey_data_using_random_intercept_factor_analysis_models.

McCrae, R. R. (2018). Method biases in single-source personality assessments. *Psychological Assessment, 30*(9), 1160–1173. https://doi.org/10.1037/pas0000566.

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences*, *44*(7), 1539–1550. https://doi.org/10.1016/j.paid.2008.01.010.

Mirowsky, J., & Ross, C. E. (1991). Eliminating defense and agreement bias from measures of the sense of control: A 2 × 2 index. *Social Psychology Quarterly, 54*(2), 127–145. https://doi.org/10.2307/2786931.

Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology, 57*(1), 401–421. https://doi.org/10.1146/annurev.psych.57.102904.190127.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrighsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322–338. https://doi.org/10.1037/a0014996.

Poropat, A. E. (2014). A meta-analysis of adult-rated child personality and academic performance in primary education. *British Journal of Educational Psychology, 84*(2), 239–252. https://doi.org/10.1111/bjep.12019.

Primi, R., De Fruyt, F., Santos, D., Antonoplis, S. & John, O. P. (2018). True or False? Keying direction and acquiescence influence the validity of socio-emotional skills items in predicting high school achievement. Submitted paper under review.

Primi, R., Santos, D., De Fruyt, F., & John, O. P. (2019). Comparison of classical and modern methods for measuring and correcting for acquiescence. *British Journal of Mathematical and Statistical Psychology.*

Primi, R., Santos, D., John, O. P., & De Fruyt, F. D. (2016). 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.

Savalei, V., & Falk, C. F. (2014a). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate Behavioral Research, 49*(5), 407–424. https://doi.org/10.1080/00273171.2014.931800.

Samuel, D. B., & Widiger, T. A. (2008). A meta-analytic review of the relationships between the five-factor model and DSM-IV-TR personality disorders: a facet level analysis. *Clinical Psychology Review, 28*(8), 1326–1342. https://doi.org/10.1016/j.cpr.2008.07.002.

Savalei, V., & Falk, C. F. (2014b). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate Behavioral Research, 49,* 407–424. https://doi.org/10.1080/00273171.2014.931800.

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.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology, 100*(2), 330–348. https://doi.org/10.1037/a0021717.

Soto, C. J., & John, O. P. (2019). Optimizing the length, width, and balance of a personality scale: How do internal characteristics affect external validity? *Psychological Assessment, 31*, 586–590.https://doi.org/10.1037/pas0000586.

Ten Berge, J. M. (1999). A legitimate case of component analysis of ipsative measures, and partialling the mean as an alternative to ipsatization. *Multivariate Behavioral Research, 34*(1), 89–102. https://doi.org/10.1207/s15327906mbr3401_4.

Valentini, F. (2017). Editorial: Influência e controle da aquiescência na análise fatorial [Editorial: Acquiescence and factor analysis]. *Avaliação Psicológica, 16,* 120–121. https://doi.org/10.15689/ap.2017.1602.

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological Methods, 15*(1), 96–110. https://doi.org/10.1037/a0018721.

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2015). The Stability of extreme response style and acquiescence over 8 years. *Assessment*. https://doi.org/10.1177/1073191115583714.

Zhang, J., & Ziegler, M. (2018). Why do personality traits predict scholastic performance? A three-wave longitudinal study. *Journal of Research in Personality, 74,* 182–193. https://doi.org/10.1016/j.jrp.2018.04.006.

Ziegler, M. (2015). "F*** You, I Won't Do What You Told Me!"—Response biases as threats to psychological assessment. *European Journal of Psychological Assessment, 31*(3), 153–158. https://doi.org/10.1027/1015-5759/a000292.

IRT Scales for Self-reported Test-Taking Motivation of Swedish Students in International Surveys

Denise Reis Costa and Hanna Eklöf

Abstract This study aims at modeling the self-reported test-taking motivation items in PISA and TIMSS Advanced studies for Swedish students using an IRT approach. In the last two cycles of the assessments, six test-specific items were included in the Swedish student questionnaires to evaluate pupil's effort, motivation and how they perceived the importance of the tests. Using a Multiple-Group Generalized Partial Credit model (MG-GPCM), we created an IRT motivation scale for each assessment. We also investigated measurement invariance for the two cycles of PISA (i.e., 2012 and 2015) and of TIMSS Advanced (i.e., 2008 and 2015). Results indicated that the proposed scales refer to unidimensional constructs and measure reliably students' motivation (Cronbach's alpha above 0.78). Differential item functioning across assessment cycles was restricted to two criteria (RMSD and DSF) and had more impact on the latent motivation scale for PISA than for TIMSS Advanced. Overall, the test-taking motivation items fit well the purpose of a diagnostic of test-taking motivation in these two surveys and the proposed scales highlighted the slight increase of pupils' motivation across the assessment cycles.

Keywords Test-taking motivation · PISA · TIMSS · IRT

1 Introduction

Regarded as a regular feature of many educational assessment systems, international surveys, such as the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS), have a major impact on the discussions about educational quality in many countries around the world (Wagemaker, 2013).

D. Reis Costa (✉)
Centre for Educational Measurement, University of Oslo,
30, 0373 Gaustadalleen, Oslo, Norway
e-mail: d.r.costa@cemo.uio.no

H. Eklöf
Department of Applied Educational Science, Umeå University, Umeå, Sweden
e-mail: hanna.eklof@umu.se

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_5

Created in 2000 by the Organisation for Economic Co-operation and Development (OECD), PISA assesses 15-year-old student's literacy in science, mathematics, and reading. First conducted in 1995 by the International Association for the Evaluation of Educational Achievement (IEA), TIMSS Advanced assesses students in the final year of secondary school enrolled in special advanced mathematics and physics programs or tracks.

As there are no personal benefits related to students' performance on the test, PISA and TIMSS Advanced are usually low-stakes tests for participating students, but high-stakes for other stakeholders. In this scenario, some pupils may lack the motivation to do their best on the test and the results, therefore, can be an underestimation of their knowledge (Eklöf and Nyroos, 2013).

In this study, we created a test-taking motivation scale for the PISA and TIMSS Advanced assessments in Sweden and we investigated the quality of these measures. These scales were built using six test-taking motivation items created specifically for each assessment and based on the expectancy-value model (Wigfield and Eccles, 2000). In particular, item response theory (IRT) analysis was used to examine the psychometric properties of the test-taking motivation items and their measurement invariance over the last two cycles of each assessment. Moreover, the differences in test motivation were studied across the different test administrations.

2 Methods

2.1 Data

A total of 4736 Swedish students participated in the PISA 2012 cycle and 5458 in 2015. In TIMSS Advanced, the 2008 assessment counted with 2303 Swedish students and in the 2015 cycle, 3937. In their questionnaires, six national items (Table 1) referring to their effort, motivation and how they perceived the importance of that specific test were presented. All items use four-point Likert-type scales and, except for negative items, they were reversed so that score categories are in increasing order with respect to the target trait, test-taking motivation.

The number of students who answered the test-taking motivation item varied by assessment and cycle. The percentage of students who had at least one missing response is larger for PISA 2012 (11%), followed by PISA 2015 (5%), and TIMSS Advanced 2008 (2%) and 2015 (1%). One possible reason for the missing data is that the student did not reach the end of the questionnaire where the motivation items were located. In PISA 2012, for example, about 80% of the students with missing responses omitted all the motivation items.

To handle this, we imputed the cases using the proportional odds model with students' background information. This analysis was conducted using the mice package (Buuren and Groothuis-Oudshoorn, 2010). As background variables, we used immigrant status, language most often spoken at home and gender for both assessments

Table 1 Item description and percentage of students with missing data by assessment and cycle

Item Code	Description	Percentage of missing response	
		PISA 2012 (%)	PISA 2015 (%)
MOTIV_R	I felt motivated to do my best on the PISA test	10.6	4.4
GODEFF_R	I engaged in good effort throughout the PISA test	10.8	4.7
DIDBES_R	I did my best on the PISA test	11.1	4.7
WORKIT_R	I worked on the tasks in the test without giving up even if some tasks felt difficult	10.9	4.5
IMPWEL_R	Doing well on the PISA test was important to me	10.7	4.7
IMP2_R	Doing well on the PISA test meant a lot to me	10.8	4.8
		TIMSS Adv. 2008 (%)	TIMSS Adv. 2015 (%)
MOTIV_R	I felt motivated to do my best on this test	2.3	1.0
DIDBE_R	I gave my best effort on this test	1.8	1.2
WORKI_R	I worked on each item in the test and persisted even when the task seemed difficult	2.0	1.1
NOCONC	I did not give this test my full attention while completing it	1.6	0.9
NOEFF	I tried less hard on this test as I do on other tests we have at school	1.9	1.8
NOWORK	While taking this test, I could have worked harder on it	1.7	2.1

Note The suffix "_R" refers to the reversing of the items. All PISA items were reversed from the original response scale: Strongly disagree (4), Disagree (3), Agree (2) and Strongly agree (1). Items were scored so that a low value is always indicative of a more negative attitude torwards the test in terms of perceived importance and reported invested effort. Likewise, three items in the TIMSS Advanced assessments were reversed: MOTIV_R, DIDBE_R and WORKI_R. In TIMSS Advanced, the original response scale was: Disagree a lot (4), Disagree (3), Agree (2) and Agree a lot (1)

and for PISA the students' economical and socio-cultural status on top of that. A sensitivity analysis comparing the IRT item parameter estimates with imputed data and with listwise case deletion was carried out and no substantial difference on the estimates and respective interval confidences were found. Thus, we proceeded the analyses using the complete cases.

2.2 Statistical Analyses

2.2.1 Descriptive Analysis

The percentage of students' agreement with the test-taking motivation items was illustrated in a radar plot. For this analysis, negative items were reversed and the response options dichotomized (with the highest value referring to the two more positive response categories of the attitudes scale).

2.2.2 Reliability

We computed the Cronbach's alpha reliability coefficient. It ranges between 0 and 1, with higher values indicating higher internal consistency of the scale. Commonly accepted cut-off values are 0.9 to signify excellent, 0.8 for good, and 0.7 for acceptable internal consistency (OECD, 2017).

2.2.3 Dimensionality

An analysis of the eigenvalues was done. Using the polychoric correlation matrix, the principal axis factor analysis and the minimum residual solution to estimate the communalities, eigenvalues were calculated using the psych package (Revelle, 2014). The eigenvalues communicate variance and guide the factor selection process by conveying whether a given factor explains a considerable portion of the total variance of the observed measures (Brown, 2014). We used the Kaiser criterion, where eigenvalues above 1.0 provide an indication of unidimensionality of the latent structure.

2.2.4 IRT Analyses

The analyses were conducted in four steps. The first step was related to the analysis of the item parameters through the Multiple-Group Generalized Partial Credit model (MG-GPCM) approach, considering each assessments cycles as a group and the estimated item parameter equal (invariant) across groups (Model 1). In the second step, an analysis of the differential item functioning (DIF) over time was carried out. For those items that did not present an indication of DIF, their parameter estimates were fixed across the groups (anchor items) in the third step of the analysis (Model 2). Finally, we estimated the individual scores for each assessment.

The MG-GPCM is based on the assumption that the two-parameter dichotomous response model governs the probability of selecting the k-th category over the (k-1) category by (Muraki, 1999):

$$P_{gjk}(\theta_g) = \frac{\exp[\sum_{r=1}^{k} Z_{gjr}(\theta_g)]}{\sum_{m=1}^{K_j} \exp[\sum_{r=1}^{k} Z_{gjr}(\theta_g)]}, \tag{1}$$

where: $Z_{gjr}(\theta_g) = Da_{gj}(\theta_g - b_{gjr})$, a_{gj} is the slope parameter for group g and item j, b_{gjr} is the item category parameter for group g, item j, and category r, D is equal to 1.7, generally inserted to make the logit scale comparable to a normal metric. The latent trait, θ_g, is generally assumed to be normally distributed for each group ($g = 1, \ldots, G$). In this study, we use the slope-intercept parameterization implemented in the mirt package (Chalmers, 2012), where $Z_{gjr}(\theta_g) = a_{gj}\theta_g + d_{gjr}$, where a_{gj} is the slope parameter for group g and item j, d_{gjr} is the intercept parameter for group g, item j, and category r.

For identification purposes, the mean and variance of the reference group (in case of PISA, the 2012 cycle and, for TIMSS Advanced, the 2008 assessment) were fixed to 0 and 1, respectively.

Since one of the advantages of the MG-GPCM approach is its flexibility to estimate item parameters separately for each group, we detected DIF using two criteria and then evaluated a second MG-GPCM model fixing anchor items. For the DIF analysis, we calculated the root mean square deviance (RMSD) for each item using the tam package (Kiefer, Robitzsch, & Wu, 2015) and the differential step functioning (DSF) using the DIFAS software (Penfield, 2005). The cut-off criteria to flag the item with DIF was an RMSD value greater than 0.3 (OECD, 2017) or large levels of DSF effect (i.e., the log-odds ratio estimator is greater than or equal to 0.64 in absolute value) as suggested in the Penfield's classification scheme (Penfield, 2008).

For the best model, individual scores were generated using weighted maximum likelihood (WLE) estimation (Warm, 1989) and were transformed to scales with a mean of 0 and a standard deviation of 1 for the reference group.

2.3 Student Weights

It is usual in these assessments to use a type of student weight (called "senate weights" in PISA) such that all countries contribute equally to the estimation of the item parameters. On PISA 2015, for example, a senate weight was constructed to sum up to the target sample size of 5000 within each country (OECD, 2017). Since the focus of this study is related only to the Swedish samples, the student weights were not included in this study.

3 Results

3.1 Descriptive Analysis

There was a significant increase in reported test-taking motivation between 2012 and 2015 in PISA. From Fig. 1, we can see a difference of more than 15 percentage points on student's agreement for all items in PISA, except on the GODEFF_R item ("I

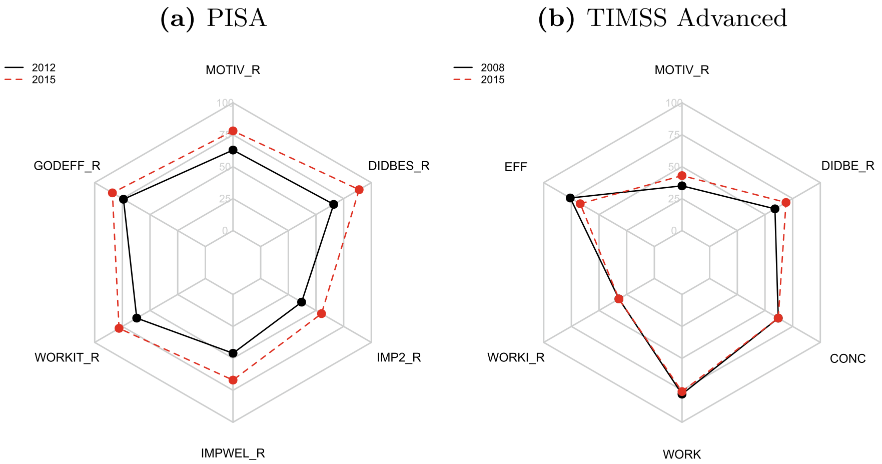


Fig. 1 Percentage of agreement by assessment and cycle. Negative items were reversed

engaged in good effort throughout the PISA test") where the difference was only 5. For TIMSS Advanced, on the other hand, the highest difference between cycles was on the DIDBE_R item ("I did my best"), where 9% of the Swedish students agree or agree a lot with this statement.

3.2 *Reliability and Dimensionality*

From Table 2, we can see that the internal consistency evaluated by Cronbach's alpha was at an acceptable level, with all values above 0.78. This measure was higher in the PISA assessments than in TIMSS Advanced. Using the Kaiser criterion, the eigenvalues suggest that the items could be adequately represented by a unidimensional scale for each assessment.

Table 2 Cronbach's α and eigenvalues by assessment and cycle

Assessment	Cycle	α	Eigenvalue 1	Eigenvalue 2
PISA	2012	0.89	4.02	0.27
	2015	0.86	3.56	0.37
TIMSS Advanced	2008	0.79	2.78	0.23
	2015	0.82	3.09	0.27

3.3 IRT Analyses

Tables 3 and 4 present the item parameter estimates through the MG-GPCM approach. By considering the item parameters invariant across the assessment cycles (Model 1), there is an improvement of half standard deviation from PISA 2012 to 2015 and about one quarter for TIMSS Advanced cycles.

Table 3 Item parameter estimates by cycle - PISA

Item code	Parameter	Model 1		Model 2	
		2012	2015	2012	2015
MOTIV_R	a1	2.304	–	2.972	2.817
	d1	2.830	–	4.265	3.309
	d2	3.713	–	0.974	0.957
	d3	0.989	–	−3.338	−3.111
GODEFF_R	a1	1.821	–	2.233	–
	d1	3.166	–	4.294	–
	d2	4.653	–	1.549	–
	d3	2.010	–	−2.994	–
IMPWEL_R	a1	2.661	–	3.239	2.957
	d1	3.109	–	4.080	3.235
	d2	2.888	–	−0.382	−0.050
	d3	−0.780	–	−4.388	−3.895
WORKIT_R	a1	1.669	–	2.145	–
	d1	2.774	–	3.752	–
	d2	3.493	–	0.798	–
	d3	1.336	–	−2.632	–
IMP2_R	a1	2.510	–	2.927	–
	d1	2.532	–	2.988	–
	d2	1.570	–	−0.994	–
	d3	−2.464	–	−4.574	–
DIDBES_R	a1	2.245	–	2.816	2.576
	d1	3.201	–	4.260	4.409
	d2	4.627	–	1.148	2.137
	d3	2.926	–	−2.505	−1.676
Group	MEAN	0	0.53	0	0.47
Group	VAR	1	0.84	1	0.79
Number of parameters		26		38	
Log-likelihood		−57971.04		−57391.65	
AIC		115994.10		114859.30	
BIC		116182.10		115134.00	

Note The symbol "–" indicates that the estimates are equal to the 2012 column

Table 4 Item parameter estimates by cycle - TIMSS Advanced

Item code	Parameter	Model 1		Model 2	
		2008	2015	2008	2015
DIDBE_R	a1	2.039	–	2.391	–
	d1	3.527	–	4.175	–
	d2	4.352	–	0.840	–
	d3	1.660	–	−3.255	–
NOCONC	a1	0.671	–	1.093	–
	d1	1.002	–	1.573	–
	d2	0.307	–	−0.808	–
	d3	−0.936	–	−2.568	–
NOEFF	a1	1.363	–	1.896	–
	d1	0.456	–	0.812	–
	d2	−0.802	–	−1.734	–
	d3	−3.332	–	−3.986	–
WORKI_R	a1	0.835	–	1.218	–
	d1	1.300	–	1.849	–
	d2	0.654	–	−0.806	–
	d3	−1.102	–	−2.928	–
MOTIV_R	a1	1.406	–	1.859	–
	d1	1.334	–	1.828	–
	d2	0.601	–	−0.984	–
	d3	−1.952	–	−3.625	–
NOWORK	a1	1.217	–	1.868	1.774
	d1	0.013	–	0.841	−0.046
	d2	−1.465	–	−1.968	−2.279
	d3	−4.237	–	−4.404	−4.559
Group	MEAN	0	0.22	0	0.27
Group	VAR	1	1.35	1	1.37
Number of parameters		26		30	
Log-likelihood		−41786.07		−41424.06	
AIC		83624.15		82908.12	
BIC		83799.36		83110.29	

Note The symbol "–" indicates that the estimates are equal to the 2012 column

Table 5 indicates good item fit for all test-motivation items for the Model 1 using the RMSD criterion. Comparing the levels of DSF effect, however, three items on the PISA dataset (MOTIV_R, IMPWEL_R, and DIDBES_R) and one item on TIMSS Advance data (NOWORK) present large DIF. Thus, we estimated the item parameters using anchor items in the analysis and freely-estimated parameters for those with large DSF (Model 2). Results indicate that flagged items were more discriminative for PISA 2012 and TIMSS Advanced 2008 than the following cycles. According to the AIC and BIC criteria, Model 2 was the best model for both assessments.

Figure 2 shows the distribution of the individual scores for Model 2 for each assessment in the logit metric with measured values ranging from −3 to 3. An increase

Table 5 The root mean square deviance (RMSD) and the item-level log-odds ratio estimate for testing the differential step functioning for each assessment

Assessm.	Item Code	RMSD		Step 1 (SE)	Step 2 (SE)	Step 3 (SE)
		2012	2015			
PISA	MOTIV_R	0.038	0.025	0.941 (0.116)	0.111 (0.068)	0.123 (0.079)
	GODEFF_R	0.043	0.027	0.522 (0.173)	0.344 (0.069)	0.385 (0.075)
	IMPWEL_R	0.031	0.024	0.709 (0.105)	−0.195(0.062)	−0.042 (0.092)
	WORKIT_R	0.041	0.029	0.098 (0.139)	−0.037 (0.060)	−0.083 (0.072)
	IMP2_R	0.033	0.016	0.275 (0.086)	0.078 (0.063)	0.013 (0.106)
	DIDBES_R	0.066	0.040	−0.172 (0.161)	−1.002 (0.079)	−0.500 (0.063)
		2008	2015			
TIMSS Adv.	DIDBE_R	0.028	0.021	−0.224 (0.132)	−0.442 (0.081)	−0.175 (0.096)
	NOCONC	0.054	0.044	0.424 (0.086)	0.278 (0.073)	−0.252 (0.105)
	NOEFF	0.036	0.019	0.324 (0.076)	−0.338 (0.088)	−0.196 (0.130)
	WORKI_R	0.042	0.034	−0.056 (0.087)	−0.314 (0.071)	−0.259 (0.110)
	MOTIV_R	0.023	0.024	−0.270 (0.088)	−0.100 (0.077)	−0.346 (0.116)
	NOWORK	0.065	0.036	0.756 (0.073)	0.189 (0.096)	0.017 (0.157)

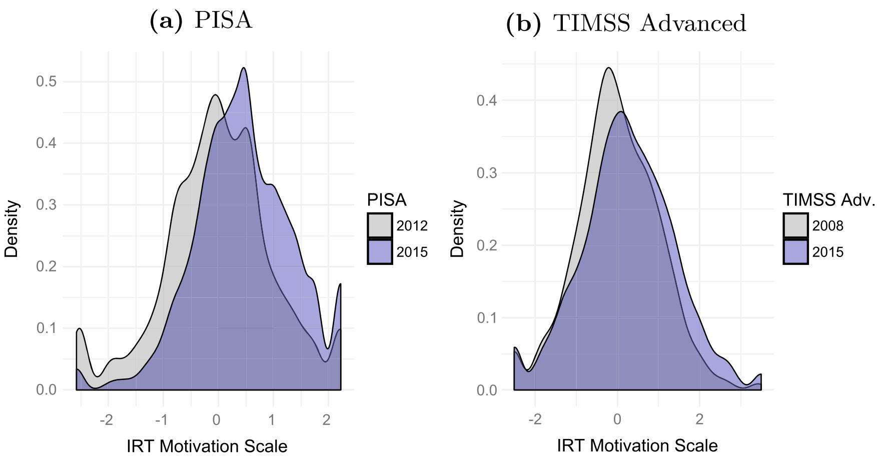


Fig. 2 Distribution of the WLE scores for Model 2 (anchor items)

in students' test motivation is observed, especially in the PISA assessment. With these measures, it is possible to monitor the test-taking motivation across assessment cycles and analyze the relationship between the reported motivation and other student measures, such as performance.

4 Discussion

In this work, we evaluated two test-taking motivation scales included in the Swedish student questionnaire of two large-scale international assessments. Results indicated that both scales are unidimensional and, while the item parameters were largely stable across the two cycles of TIMSS Advanced, half of the items in PISA showed some DIF. Our findings also indicated that there was a slight increase in test-taking motivation in PISA 2015 in comparison to 2012 which can be related to the change of the test mode administration from paper and pencil to computer-based across these two PISA cycles.

For future studies, we intend to expand these analyses to carry out studies of PISA log-file data to evaluate how the self-reported test-taking motivation measures agree with student behaviors during the administration of the test. With information of response times, for example, we can further investigate the construct validity of these measures.

References

Brown, T. A. (2014). *Confirmatory factor analysis for applied research*. New York: Guilford.

Buuren, S. V., & Groothuis-Oudshoorn, K. (2010). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–68.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

Eklöf, H., & Nyroos, M. (2013). Pupil perceptions of national tests in science: Perceived importance, invested effort, and test anxiety. *European Journal of Psychology of Education*, *28*(2), 497–510.

Kiefer, T., Robitzsch, A., & Wu, M. (2015). Tam: Test analysis modules. *R Package*.

Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple-group partial credit model. *Journal of Educational Measurement*, *36*(3), 217–232.

OECD. (2017). *PISA 2015 Technical Report*. Paris: OECD Publishing. https://doi.org/10.1787/9789264255425-en.

Penfield, R. D. (2005). Difas: Differential item functioning analysis system. *Applied Psychological Measurement*, *29*(2), 150–151.

Penfield, R. D. (2008). Three classes of nonparametric differential step functioning effect estimators. *Applied Psychological Measurement*, *32*(6), 480–501.

Revelle, W. (2014). Psych: Procedures for psychological, psychometric, and personality research. *R Package*.

Wagemaker, H. (2013). International large-scale assessments: From research to policy. In L. Rutkowski, M. von Davier, & D. Rutkowski D (Eds.), *Handbook of international large-scale* *assessment: Background, technical issues, and methods of data analysis* (pp. 11–35). New York: Chapman Hall/CRC.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81.

A Modification of the IRT-Based Standard Setting Method

Pilar Rodríguez and Mario Luzardo

Abstract We present a modification of the IRT-based standard setting method proposed by García, Abad, Olea & Aguado (Psicothema 25(2):238–244, 2013), which we have combined with the cloud delphi method (Yang, Zeng, & Zhang in IJUFKBS 20(1):77–97, 2012). García et al. (Psicothema 25(2):238–244, 2013) calculate the average characteristic curve of each level, to determine cutoff scores on the basis of the joint characteristic curve. In the proposed new method, the influence of each item on the average item characteristic curve is weighted according to its proximity to the next level. Performance levels are placed on a continuous scale, with each judge asked to determine an interval for each item. The cloud delphi method is used until a stable final interval is achieved. From these judgments, the weights of each item in the scale are calculated. Then, a family of weighted average characteristic curves is calculated and in the next step, joint weighted averaged ICC are calculated. The cutoff score is determined by finding the ability where the joint weighted averaged ICC reach a certain predefined probability level. This paper compares the performance of this new procedure for a math test with the classic Bookmarking method. We will show that this modification to the method improves cutoff score estimation.

Keywords Performance standard setting · Item response theory · Delphi method

1 Introduction

The setting of performance standards is a central issue in educational measurement. Therefore, the methods for setting them have undergone significant development in recent years. It has been one of the most researched topics over the last forty years

P. Rodríguez (✉) · M. Luzardo
Eastern Regional University Center, University of Republic, Maldonado, Uruguay
e-mail: prodriguez@cure.edu.uy

M. Luzardo
e-mail: mluzardo@psico.edu.uy

M. Luzardo
School of Psychology, University of Republic, Montevideo, Uruguay

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_6

and also one of the most contentious (Berk, 1986; Cizek & Bunch, 2007; Glass, 1978; Hambleton, 1978; Hambleton et al., 2000; Jaeger, 1989; Kane, 1994; Linn, 2003; Margolis & Clauser, 2014; Mousavi, Cui & Rogers, 2018). Different methods of setting cutoff scores provide different standards on the same test (Jaeger, 1989). Therefore, it is important to develop methods to set cutoff scores with precision and stability. This work is a contribution in such regard.

2 Method

We present a procedure which introduces a modification to the method for establishing cutoff scores devised by García et al. (2013), combined with the cloud delphi method. The proposed method can be applied to both a bank already built, and bank items built to match a certain performance level.

Let us assume that *k* levels of performance have been defined (for instance, level 1, level 2 and level 3).

In García et al. (2013)'s method, the bank is built to obtain a set of items that will represent each performance level; but it cannot be agreed that all the items classified or developed for each performance level represent the description of that level in the same way.

To capture the difference in the influence of each item for the determination of the cutoff scores, we resort to the cloud delphi method. To apply this method, it is necessary to obtain a continuous magnitude of the performance level of each item. Operationally, a correspondence of the levels is established with the interval [0, k + 1], with the integer values 1, 2, …, k being the lower ends of the levels expressed qualitatively. For instance, if there are three performance levels, the interval will be (0, 4), with the sub-interval (0, 1) corresponding to "does not reach level 1," interval [1, 2) to level 1, interval [2, 3) to level 2, and (3, 4) to level 3. There is a bijective function between the scale of skill and performance levels.

From a group of judges and by applying the cloud delphi method, a numerical value is obtained on the performance scale: where the item will have a subjective probability of 0.5 of being correctly responded to by a subject with that value on the scale. It is a difficult task for a judge to determine the point of the scale where the above property is fulfilled. However, the proposed method asks each judge to determine an interval on the performance scale, where he considers a subject with that performance level to have a 0.5 probability of correctly responding to the item. The width of the interval will reflect the uncertainty in the judge's response. The cloud delphi method allows us to stabilize their response, and the intervals provided by each judge can be used to determine the item score on the performance scale. This value of each item determines a position on that scale, which will then be used to weight its influence on the establishment of each cutoff score.

After the items have been calibrated, the ICC of each item can be used to calculate the weighted ICC in relation to each cutoff score, which we will note as $WP_k(\theta)$.This curve connects the performance level scale with the ability scale, and represents the probability that a subject will correctly respond to a typical item of cutoff score *k*.

From the $WP_k(\theta)$ we can find the joint probability of correctly responding to a prototype item of cutoff score k and the previous cutoff scores. We will note this curve as $JWP_k(\theta)$.The cutoff score will be determined as the value of the ability that causes the joint probability to reach a predetermined value μ (for instance, 0.5); that is, it solves the equation $JWP_k(\theta) = \mu$.

2.1 Cloud Delphi Method

The cloud model relates a qualitative concept with quantitative data based on probability and the fuzzy set theory. The most important model here is the normal cloud model, based on the normal distribution and the Gaussian membership function. In particular, the normal cloud model makes it possible to measure the deviation of a random phenomenon from a normal distribution, when the former does not strictly satisfy the latter (Wang, Xu, & Li, 2014).

This model uses three numerical concepts: expectation (*Ex*); entropy (*En*), which represents the degree of cloudiness of the concept; and hyper entropy (*He*), which represents the variability of the concept in the cloud (Yang, Zeng, & Zhang, 2012).

Formally, let us denote U as the universe of discourse, which is made up of numbers, and let *T* be a qualitative concept. Let us assume that concept *T* is determined in U by its expectation, entropy and hyper entropy; in other words, by the triple *(Ex, En, He).* Let $x \in U$ be a random realization of concept *T,* such that *x* has normal distribution of mean *Ex* and variance σ_x^2. In addition, we assume that σ_x^2 is a random variable with a normal distribution of mean *En* and variance He^2. Let $\mu_T(x) \in [0, 1]$ be the certainty degree of *x* belonging to *T*. We will say that the distribution of *x* over U is a normal cloud if

$$\mu(x) = e^{\frac{(x-Ex)^2}{2(y)^2}} \; with \, y \sim N\left(En, He^2\right) \tag{1}$$

Then, the distribution of x in universe U is defined as a cloud and x is called cloud drop. This definition establishes that drop $x \in U$ is an extension of concept T. Mapping $\mu_T(x)$ establishes that the certainty degree of x belonging to concept T is a probability distribution (Yang et al., 2012).

The procedure for applying the cloud delphi method was developed by Yang et al. (2012), and we applied it by following the procedure explained in the previous section to obtain the level of each item.

A set of *n* judges was asked to determine the interval on the performance level scale in which they think a subject has a 0.5 probability of correctly responding to the item. The procedure involves the following steps:

Step 1 : Set the iteration counter *j* equal to one.

Step 2 : In iteration *j*, each judge provides the requested interval. The following intervals are thus obtained $\left[l_i^{(j)}, u_i^{(j)}\right]$, where *i* indicas the *i*-th judge.

Step 3 : The interval provided by each judge is expressed in terms of the normal cloud model, determined by the triple $C_i^{(j)} = \left(Ex_i^{(j)}, En_i^{(j)}, He_i^{(j)}\right) i = 1, \ldots, n$.

Cloud parameters can be calculated as follows for *i* = 1,…, *n*:

$$\begin{gathered} Ex_i^{(j)} = \frac{l_i^{(j)}+u_i^{(j)}}{2} \\ En_i^{(j)} = \frac{u_i^{(j)}-l_i^{(j)}}{6} \\ He_i^{(j)} = \frac{\max\{u_i^{(j)}-u_i^{(j-1)},0\}+\max\{l_i^{(j-1)}-l_i^{(j)},0\}}{6} \; and \; He_i^{(1)} = \frac{En_i^{(1)}}{6} \end{gathered} \tag{2}$$

Step 4 : Generate the feedback information for the next iteration by using cloud aggregation algorithms described by Yang et al. (2012).

The synthetic cloud and weighted cloud of each item are shown graphically: to each judge for the purpose of making a new estimate of the interval. These clouds are determined by means of the following equations:

Synthetic Cloud

Let us assume we have *n* clouds $C_i = (Ex_i, En_i, He_i) i = 1, \ldots, n$. The parameters of synthetic cloud $C_s(Ex_s, En_s, He_s)$ are defined by

$$\begin{gathered} Ex_s = \frac{1}{n}\sum_{i=1}^{n} Ex_i \\ En_s = \frac{1}{6}\left[\max_i\{Ex_i + 3En_i\} - \min_i\{Ex_i - 3En_i\}\right] \\ He_s = \frac{1}{n}\sum_{i=1}^{n} He_i \end{gathered} \tag{3}$$

Weighted Cloud

The parameters of weighted cloud $C_{wa}(Ex_{wa}, En_{wa}, He_{wa})$ are defined by

$$\begin{gathered} Ex_{wa} = \sum_{i=1}^{n} w_i Ex_i \\ En_{wa} = \sqrt{\sum_{i=1}^{n} (w_i En_i)^2} \\ He_{wa} = \sqrt{\sum_{i=1}^{n} (w_i He_i)^2} \end{gathered} \tag{4}$$

The relative importance of each judge in the *j-th* step is:

$$r_i^{(j)} = \frac{1}{\left|\frac{\left(Ex_i^{(j)} - Ex_s^{(j)}\right)}{Ex_s^{(j)}}\right| + En_i^{(j)} + He_i^{(j)}} \quad i = 1, 2, \ldots, n \tag{5}$$

Finally, the weights are:

$$w_i^{(j)} = \frac{r_i^{(j)}}{\sum_{i=1}^{n} r_i^{(j)}} \quad i = 1, 2, \ldots, n \tag{6}$$

Step 5 : The relative difference of the entropy with respect to the previous iteration, which we will denote as ΔEn; and the *Unc* ratio of hyper entropy with respect to the entropy, are calculated for the *j-th* iteration.

$$\begin{gathered}\Delta En_i^{(j)} = \frac{\left|En_i^{(j-1)} - En_i^{(j)}\right|}{En_i^{(j-1)}} \quad y \; \Delta En_i^{(1)} = En_i^{(1)} \\ Unc_i^{(j)} = \frac{He_i^{(j)}}{En_i^{(j)}} i = 1, \ldots, n\end{gathered} \tag{7}$$

Step 6 : If $Unc_i^{(j)} = 0$ and for δ > 0 prefixed $\Delta En_i^{(j)} \leq \delta \quad i = 1, 2, \ldots, n$ iterations are completed.

The cloud delphi method is applied until the opinion stabilizes. Once the final intervals have been obtained, the synthetic cloud and weighted cloud are obtained. The weighted cloud of each item is considered the final decision of the judges; and its expectation, which we will denote as b_i, will be the score of the item on the performance scale.

To illustrate the information received by a judge, Fig. 1 shows the graph for item 230 of a mathematics test.

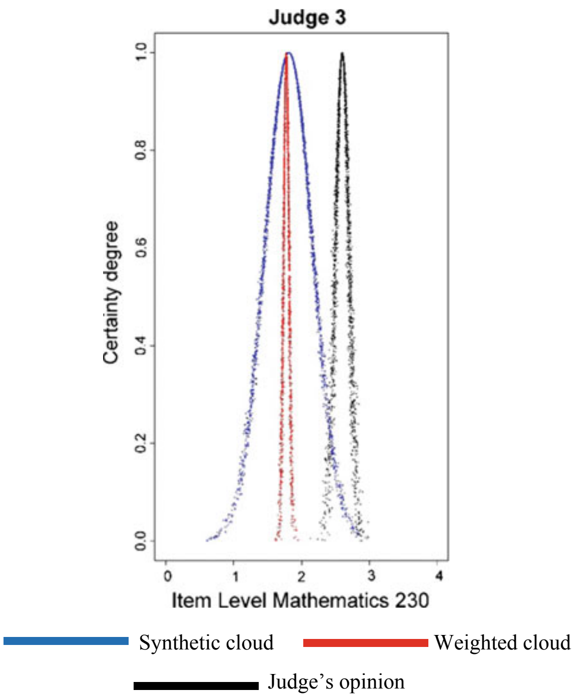
2.2 Setting Cutoff Scores

This second stage involves generalizing García et al. (2013)'s method to obtain the cutoff scores. From the ICCs of each item, the weighted average ICC is obtained at each cutoff score.

$$WP_k(\theta) = \frac{\sum_{j=1}^{N} K\left(\frac{b_j - k}{h}\right) P_j(\theta)}{\sum_{j=1}^{N} K\left(\frac{b_j - k}{h}\right)} \tag{8}$$

where b_j is the item score estimation on the performance scale, h is the bandwidth and K is a kernel. Kernels are used to determine the weight of each item in the weighted average ICC estimation.

Fig. 1 Cloud model showing synthetic cloud, weighted cloud and judge's opinion for mathematics item 230



Joint averaged ICC (JWP) is calculated for the cutoff scores and represents the probability that an examinee with ability θ will respond correctly to the prototype item of cutoff score k and all previous ones. It is calculated by means of $JWP_k(\theta) = \prod_{z=1}^{k} WP_z(\theta)$.

To calculate cutoff score k, we identify the ability for which the probability of responding to the prototype item of cutoff score k and the previous ones is equal to a predetermined value. We denote probability with μ and the examinee's ability with θ_{cs}. This ability is the solution to the equation $JWP_k(\theta_{cs}) = \mu$, from which the cutoff score is obtained.

3 Results

The method was tested in a university entrance exam assessing reading and mathematics (Rodríguez, 2017). Two methods were applied for performance standard setting: bookmark and the method proposed in this paper.

The item bank has 247 items; a sample of 50 reading and 50 mathematical items was taken. Judges established three performance levels. For the proposed method, two kernels were applied: Gaussian and Epanechnikov. The Gaussian kernel is defined by $\frac{1}{\sqrt{2\pi}}e^{-u^2/2}$; the Epanechnikov kernel by $\frac{3}{4}\left(1 - u^2\right)$, with $|u| \leq 1$. Results for different methods are shown in Tables 1, 2, 3 and 4.

Reading

The final cutoff scores from the proposed method represent the average of the results in both kernels.

Mathematics

For the Mathematics test, the final cutoff scores by the proposed method represent the average of the results in both kernels.

A sample was selected of 204 students who took the exam. Their performance levels were classified using the bookmark method and proposed method. They were also classified by expert judgment. The proportions of students in each level are presented in the following graph (Fig. 2).

Table 1 Cutoff scores obtained by Bookmark method in the three performance levels for the Reading test

Levels	Bookmark
1	−0.94
2	0.12
3	1.62

Table 2 Cutoff scores obtained by the proposed method using Gaussian and Epanechnikov kernels in the three performance levels for the Reading test

Levels	Epanechnikov	Gaussian	Average
1	−1.63	−1.47	−1.55
2	−0.19	−0.07	−0.13
3	1.27	1.37	1.32

Table 3 Cutoff scores obtained by Bookmark method in the three performance levels for the Mathematics test

Levels	Bookmark
1	−1.23
2	−0.09
3	1.57

Table 4 Cutoff scores obtained by the proposed method using Gaussian and Epanechnikov kernels in the three performance levels for the Mathematics test

Levels	Epanechnikov	Gaussian	Average
1	−0.88	−0.94	−0.91
2	−0.01	0.13	0.06
3	1.39	1.43	1.41

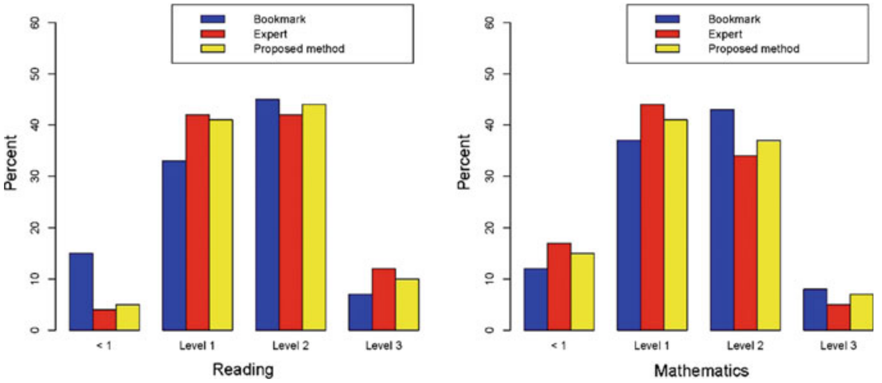


Fig. 2 Comparative graphs of the two methods and expert judgment for the cutoff scores of each level of the Mathematics and Reading tests

4 Discussion

The proposed method establishes cutoff scores closer to the expert judgment than the bookmark method. Moreover, it is better at capturing the variability of the item bank and manages to weight the qualitative judgments. It differs from the bookmark method in that all items participate in determining the cutoff score beyond its ordering by level of difficulty. In addition, it avoids the confusion and discrepancies of the bookmark method when there is no agreement between the difficulty obtained through the theory of response to the item, and a judge's perception of subjective difficulty related to the item. This method considers both the empirical difficulty and the judges' relative difficulty, with both participating in determining the cutoff scores.

This method also allows greater variability in the judges' opinion, capturing the fuzziness of the process; it does not require the determination of a score, but an interval, which makes the task simpler and more efficient.

Unlike García et al. (2013)'s original method, which requires that the items are developed for a certain performance level, it can be applied to banks of previously developed items. The proposed method is also more flexible, as the original considers the items developed for each level to contribute with the same magnitude to each cutoff score. Therefore, this approach makes it possible to obtain a more adjusted valuation of the contributions of each item in the continuum representing the performance level. These advantages make the proposed method a better alternative for the establishment of cutoff scores.

References

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion. *Review of Educational Research, 56*(1), 137–172.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting. A guide to establishing and evaluating performance standards on tests*. Thousand Oak, CA: Sage Publications.

García, P. E., Abad, F. J., Olea, J., & Aguado, D. (2013). A new IRT-based standard setting method: Application to elath-listening. *Psicothema, 25*(2), 238–244.

Glass, G. (1978). Standards and criteria. *Journal of Educational Measurement, 15*(4), 237–261.

Hambleton, R. K. (1978). The use of cut-off scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement, 15*(4), 277–290.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24*(4), 355–366.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (pp. 485–514). New York: American Council on Education and Macmillan.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425–461.

Linn, R. (2003). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives, 11*(31). Retrieved from: http://epaa.asu.edu/epaa/v11n31/.

Margolis, M. J., & Clauser, B. E. (2014). The impact of examinee performance information on judges' cut scores in modified Angoff standard setting exercises. *Educational Measurement Issues and Practice, 33*(1), 15–22.

Mousavi, A., Cui, Y., & Rogers, T. (2018). An examination of different methods of setting cutoff values in person fit research. *International Journal of Testing*, 1–22. https://doi.org/10.1080/15305058.2018.1464010.

Rodríguez, P. (2017). Creación, desarrollo y resultados de la aplicación de pruebas de evaluación basadas en estándares para diagnosticar competencias en matemática y lectura al ingreso a la universidad. *Revista Iberoamericana de Evaluación Educativa, 10*(1), 89–107. https://doi.org/10.15366/riee2017.10.1.005.

Wang, G., Xu, C., & Li, D. (2014). Generic normal cloud model. *Information Sciences, 280,* 1–15.

Yang, X. J., Zeng, L., & Zhang, R. (2012). Cloud delphi method. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 20*(1), 77–97.

Model Selection for Monotonic Polynomial Item Response Models

Carl F. Falk

Abstract One flexible approach for item response modeling involves use of a monotonic polynomial in place of the linear predictor for commonly used parametric item response models. Since polynomial order may vary across items, model selection can be difficult. For polynomial orders greater than one, the number of possible order combinations increases exponentially with test length. I reframe this issue as a combinatorial optimization problem and apply an algorithm known as simulated annealing to aid in finding a suitable model. Simulated annealing resembles Metropolis-Hastings: A random perturbation of polynomial order for some item is generated and acceptance depends on the change in model fit and the current algorithm state. Simulations suggest that this approach is often a feasible way to select a better fitting model.

Keywords Combinatorial optimization · Nonparametric item response theory · Monotonic polynomial · Balanced incomplete block design

Many standard unidimensional item response models assume a normally distributed latent trait and a simplistic relationship between the latent trait and the item responses. For example, the two-parameter logistic model (2PL) represents a multivariate extension of logistic regression, where the log-odds of obtaining a correct response to the items is a linear function of the latent trait (Birnbaum, 1968). This relationship may not be expected to hold for all educational and psychological constructs (Meijer & Baneke, 2004), and violations may arise from population heterogeneity in exposure to unique item content (Falk & Cai, 2016b) or items that require multiple steps in order to complete (Lee & Bolt, 2018). Additional flexibility in the trait-response relationship is possible, including but not limited to nonparametric Kernel smoothing (Ramsay, 1991), smoothed isotonic regression (Lee, 2007), Bayesian nonparametric techniques (Duncan & MacEachern, 2013), normal ogive models that assume

C. F. Falk (✉)
Department of Psychology, McGill College, McGill University, 7th Floor,
Montreal, QC 2001, H3A 1G1, Canada
e-mail: carl.falk@mcgill.ca
URL: https://www.mcgill.ca/psychology/carl-f-falk

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_7

heteroscedastic errors (Molenaar, 2015), and splines (Ramsay & Wiberg, 2017). Alternatively, if the source of this assumption violation stems in part from a non-normal trait distribution, one could directly model such non-normality (Woods, 2007).

The focus of this paper is on a monotonic polynomial (MP) approach to flexible item response function (IRF) estimation (Falk & Cai, 2016a, 2016b; Liang & Browne, 2015). The basic idea behind MP item response models is to replace the linear predictor of a standard item response model with an MP. Like nonparametric techniques, MP models make few assumptions about the underlying process that produces non-standard response functions. Rather, increasing polynomial order allows MP models to approximate many different functional forms, regardless of whether the MP is the true model (Feuerstahler, 2016). In contrast to the 2PL, a logistic function of a monotonic polynomial models the log-odds of a correct response as a polynomial function of the latent trait with constraints imposed such that this relationship is monotonic.

We believe the MP approach warrants further study for its potential to fulfill several needs of large scale or operational testing. For example, a psychometrician may use an MP-based model to improve item fit for a few items on a long test, allowing retention of expensive-to-develop items, but still use a traditional item model such as the 2PL or three-parameter logistic (3PL) for the remaining test items. Since MP-based models can also be explained using an analogy with polynomial regression, MP-based approaches may be more substantively interpretable to some stakeholders. We also conjecture that the derivatives necessary for MP-based item models to be used in a computer adaptive test with traditional item selection strategies are readily available in closed form, in contrast to some other approaches (Xu & Douglas, 2006). Finally, a testing program that has hundreds of items is likely to employ a planned missing data design. It would otherwise be burdensome to expect respondents to complete all such test items in a diligent manner. MP-based item models can be used in conjunction with maximum marginal likelihood (MML) estimation (Bock & Aitkin, 1981), which can be used with planned missing data designs and investigations of differential item functioning (Falk & Cai, 2016a).

1 The Computational Problem

One potential barrier for MP-based models involves a computational problem in selecting polynomial order. To further understand, consider the IRF for a logistic function of a monotonic polynomial (Falk & Cai, 2016a; Liang & Browne, 2015):

$$P_j(1|\theta) = \frac{1}{1 + \exp(-(c_j + m_j(\theta)))} \tag{1}$$

where $j = 1, \ldots, n$ indexes n test items, θ corresponds to the latent trait, and $m_j(\theta)$ is a polynomial term:

$$m_j(\theta) = b_{1,j}\theta + b_{2,j}\theta^2 + \cdots + b_{2k_j+1,j}\theta^{2k_j+1} \tag{2}$$

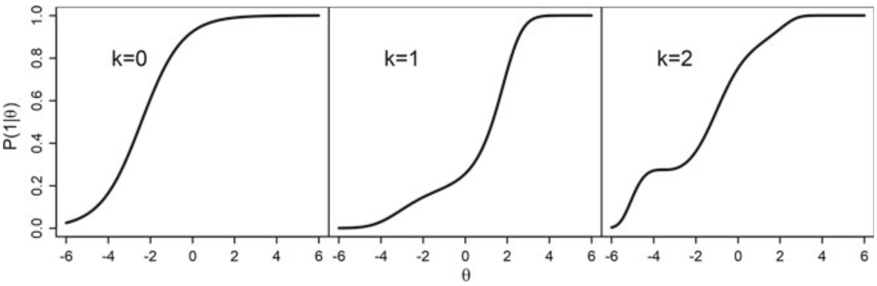


Fig. 1 Example response functions for three different order polynomials

$m_j(\theta)$ is parameterized to be monotonic increasing and has a non-negative derivative with respect to θ. This is accomplished in part by a polynomial with an odd number of terms: $2k_j + 1$, where k_j is a non-negative integer that controls polynomial order for item j (see Fig. 1). In addition, the coefficients, $b_{1,j}, b_{2,j}, \ldots, b_{2k_j+1,j}$, are not directly estimated, but are a function of $2k_j + 1$ other parameters with constraints that maintain monotonicity. Other MP models have been developed based on the 3PL, generalized partial credit, and graded response models (Falk, 2018; Falk & Cai, 2016a, 2016b). When $k_j = 0$, these models reduce to their standard counterparts (e.g., Eq. 1 reduces to the 2PL).

The key to the computational problem concerns the selection of k_j, which may be different for each item. This problem is a byproduct of using MML for estimation: Selection of k_j for one item may affect item fit for other items and overall model fit. In one investigation, Falk and Cai (2016a) employed a step-wise approach whereby AIC was used to select a single increase in polynomial order for one item at a time. This approach is difficult to use with a long test as each step would require fitting n models. For example, if $n = 100$, then 100 models must be fit before increasing polynomial order for a single item. In a different paper, Falk and Cai (2016b) experimented with use of summed score item fit statistics, $S - X^2$ (Orlando & Thissen, 2000), to screen for items that may be good candidates for use of an MP. Although this approach arguably improved fit, $S - X^2$ had power that was less than expected to detect non-standard items, and using summed score based item fit statistics may not always be desirable with missing data. If an observed score substitute for θ is used in estimation instead, then the modeler may proceed item by item in selection of k_j. However, this approach may not readily handle multiple group models or models with missing data.

1.1 A Possible Solution

We reframe the selection of k_j for each item as a combinatorial optimization problem. If we consider k_j for each item from 0 to 2, then there are 3^n possible combinations of polynomial order. Clearly for large n, there may be many combinations and therefore too many possible models to actually try out even with a modern computer. Further suppose that there is some combination of polynomial order that may be optimal (e.g., according to information criterion such as AIC or BIC). In addition to a stepwise approach being computationally slow, it may also be prone to getting stuck at a local optimum.

Although there are a number of combinatorial optimization algorithms suitable for finding an approximate global optimum, we chose to experiment with simulated annealing (Černý, 1985; Kirkpatrick, Gelatt, & Vecchi, 1983), which has seen some use by psychometricians (Drezner & Marcoulides, 1999; Edwards, Flora, & Thissen, 2012). SA gets its name in part from an analogy to metallurgy, yet we find it more intuitive to explain its workings by analogy to Metropolis-Hastings (MH). Given some model, M_s, at iteration s, SA has the following steps:

1. Generate some candidate model, M_s^*, from a list of possible neighboring models in a well-defined search space.
2. Compute *energy* for the candidate, $e(M_s^*)$, and current model, $e(M_s)$.
3. Determine acceptance/rejection of the candidate, M_s^*, based on the difference in energy, $e(M_s^*) - e(M_s)$, and the current *temperature*, t_s, which represents the current algorithm state.
4. Repeat 1–3 and stop based on an iteration maximum, S, or termination rule.

In the same way that MH will tend to move towards and sample from high-density regions of a probability distribution, SA will tend to move towards and select models in regions of a search space that have better fit. In our application, we allowed values for k_j between 0 and 2, which defines the search space as the 3^n possible polynomial order combinations. We considered a neighboring model to be a random increment or decrement of 1 to k_j for one or two items that were randomly sampled with uniform probability. For example, if item 5 were to be randomly selected and the current $k_5 = 1$, then the candidate could only change to $k_5 = 0$ or $k_5 = 2$ (selected with equal probability). If $k_5 = 0$, then the candidate had $k_5 = 1$. k_j for all other items would remain as-is. Changing k_j by only one at a time for each item and only for a couple of items may allow a reduction in estimation difficulty. For example, use of parameter estimates from a lower-order polynomial may be used as starting values for some parameters when estimating models with higher-order polynomials. However, defining neighbors and the search space in this way, it is possible to move from one state of the search space (e.g., all $k_j = 0$) to the furthest state (e.g., all $k_j = 2$) within only 300 or 150 steps or less if $n = 100$ and either one or two items' polynomials are perturbed at each step.

Energy is a function of the fitted model and defines its optimality. For this purpose, we used $e(\cdot)$ to calculate either AIC or BIC. Thus, lower energy indicates better fit. The acceptance probability of M_s^* was based on the following,

$$\min\left\{1, \exp(-(e^* - e)/t_s)\right\} \tag{3}$$

where we use e^* and e as shorthand for $e(M_s^*)$ and $e(M_s)$, respectively. In other words, if M_s^* has lower energy (or improves model fit), it is accepted with certainty. If M_s^* results in worse fit, the model may still be accepted with some non-zero probability. The function in (3) is based on Kirkpatrick and colleagues' work (Kirkpatrick et al., 1983) and is often used in applications of SA, in part due to its similarity to acceptance probabilities under MH (see p. 672).

Acceptance of a suboptimal model may still be useful, especially early in the algorithm, to the extent that it allows SA to avoid being stuck in a local optimum. However, t_s typically decreases across iterations as determined by a cooling schedule so that the probability of accepting a suboptimal model is less likely over time. A conceptual explanation of this behavior is as follows. If after many iterations SA has led M_s to (hopefully) be near the global optimum, a lower value for t_s will provide increasingly smaller acceptance probabilities for suboptimal models, potentially forcing additional accepted models to be closer and closer to the optimum.

Although there is a rich literature on the selection of a starting value and cooling schedule for t_s, in this paper we opted for a simplistic solution as a preliminary test of SA's potential. In particular, we considered starting temperatures of 5, 10, and 25. To provide a concrete example, suppose an increase in BIC of 10 is very undesirable. With $t_s = 5$, $t_s = 10$, and $t_s = 25$ such an increase would yield acceptance of approximately .14, .37 and .67, respectively, meaning that in most cases such a model would be accepted when $t_s = 25$, but rejected when $t_s = 5$ or $t_s = 10$. We chose a straight cooling schedule in which temperature decreases linearly across iterations: $t_s = t_0(S - s)/S$, where t_0 is the starting temperature. Though we note that finer tuning may result in slightly better performance (Stander & Silverman, 1994).

2 Simulations

Simulations were conducted to test the ability of SA to select polynomial order for MP-based item models. The main outcome was item response function recovery, followed by whether SA correctly modeled non-standard items with an MP. A final purpose was to test MP-based models along with SA under conditions that might occur with a planned missing data design.

2.1 Method

Fixed Factors. Simulated datasets included 100 dichotomous items, 5000 respondents, and a standard normal θ. Twenty-five replications per cell of the below data generation design were conducted, with data generation in R (R Core Team, 2015) and models fitted using *rpf* Pritikin (2016) and *OpenMx* Neale et al. (2016).

Data Generation. We manipulated the percentage of items that followed a non-standard model (20, 40, 60, and 80%), with such IRFs generated as the cumulative distribution function (CDF) from a mixture of normal variates, $p_1\mathcal{N}(\mu_1, \sigma_1^2) + p_2\mathcal{N}(\mu_2, \sigma_2^2) + p_3\mathcal{N}(\mu_3, \sigma_3^2)$. To generate variety in IRFs across items and datasets, the following values were randomly generated, $p_1 \sim \text{unif}(.1, .6)$, $p_2 \sim \text{unif}(.1, .3)$, $p_3 = 1 - p_1 - p_2$, $\mu_1 \sim \mathcal{N}(-2.2, .2^2)$, $\mu_2 \sim \mathcal{N}(2.2, .2^2)$, $\mu_3 \sim \mathcal{N}(0, .2^2)$, $\sigma_1 \sim \mathcal{N}(2, .3^2)$, $\sigma_2 \sim \mathcal{N}(.6, .3^2)$, and $\sigma_3 \sim \mathcal{N}(.6, .3^2)$. The remaining items followed a normal CDF (analogous to a normal ogive model) with $\mu \sim \text{unif}(-2.5, 2.5)$ and $\sigma \sim \mathcal{N}(2, .4^2)$. Although the MP-based item model does not strictly follow the exact same shape as the normal CDF items, we still consider them "standard" items for the following investigation since these items should be well approximated by a 2PL or MP with $k = 0$.

We also compared complete data (all respondents completed all items) versus missing data. The missing data condition emulated a planned missing data design with a balanced incomplete block design. The 100 items were split into 5 blocks of 20 items each. Ten test forms with 40 items (i.e., 2 blocks) each were created, corresponding to 60% missing data. We argue that this number of items per test-taker is not atypical of such a design for item calibration, while 60% missing data at this sample size may pose a challenge for MP models.

Fitted Models. To all datasets, we used the logistic function of an MP in (1) as parameterized by Falk and Cai (2016a) and included three models in which k was fixed to the same for all items: $k = 0$, $k = 1$, and $k = 2$. Note that $k = 0$ corresponds to the 2PL model. Following these models, we performed several runs of SA by crossing the following conditions: Energy (AIC vs. BIC), starting temperature (5, 10, and 25), and number of items to perturb (1 vs. 2). One of the three fixed models with the best energy was chosen as the starting model for each SA run. For all MP models with $k > 0$, soft Bayesian priors following Falk and Cai (2016a) were used. One additional model followed the same procedure as SA and started at the best fixed k model according to BIC, but all candidate models were accepted with certainty. We refer to this approach as *semi-random* in what follows, and was included to test whether SA has any advantage over a completely random search in the neighborhood of the best BIC of the fixed models. This and all SA runs included only 300 iterations, and the best model according to AIC or BIC was recorded as the selected model, regardless of whether it was the last accepted model.

2.2 *Results*

Response Function Recovery. Recovery of response functions was assessed using root integrated mean square error (RIMSE) (Ramsay, 1991), using $Q = 101$ equally spaced quadrature points (X_q) between -5 and 5:

$$\text{RIMSE}_j = \left(\frac{\sum_{q=1}^{Q} ((\hat{P}_j(1|X_q) - P_j(1|X_q))^2 \phi(X_q)}{\sum_{q=1}^{Q} \phi(X_q)} \right)^{1/2} \times 100 \qquad (4)$$

which can be understood as the root of a weighted mean of squared differences between true, $P_j(1|X_q)$, and estimated, $\hat{P}_j(1|X_q)$, response functions, with the population density for θ, $\phi(X_q)$, providing weights. Lower values of RIMSE are better, and the values we report were averaged across all items and replications in each cell of the simulation design.

In general, differences across most tuning options for SA were small for RIMSE, with the number of item perturbations and starting temperature resulting in differences in average RIMSE less than .1 in each cell of the data generation design. For brevity, we report RIMSE results using an initial temperature of 5 and a single item perturbation per iteration of the algorithm. This starting temperature slightly outperformed the other SA conditions.

The best (according to AIC or BIC) out of the fixed (all $k = 0$, $k = 1$, or $k = 2$) models was compared with the 2PL (all $k = 0$), SA (using AIC or BIC), and semi-random models, and RIMSE for these models appears in Table 1. The best performing model is highlighted in bold for each column, and the second best in bold and italics. We highlight several noticeable trends. First, AIC tended to do better than BIC with complete data and a higher percentage of non-standard items. This result tended to hold regardless of whether SA or a fixed k was utilized. For example, with complete data and 80% non-standard items, use of AIC resulted in RIMSE of 1.98 and 1.95, for SA and fixed conditions, respectively, whereas BIC resulted in 2.42 and 2.91. With only 20% non-standard items, BIC performed better than AIC, and this was especially true under missing data where SA using AIC (RIMSE = 1.93) had worse performance than SA with BIC (RIMSE = 1.77) and all other models (RIMSE = 1.81). SA in conjunction with AIC selection was otherwise the best or second best performing model across all other conditions. However, we note that SA with BIC *always* outperformed the 2PL and semi-random conditions. In contrast, SA with

Table 1 Root integrated mean square error (response function recovery)

	Complete data				Missing data			
Model	20%	40%	60%	80%	20%	40%	60%	80%
SA (AIC)	***1.45***	**1.57**	**1.81**	***1.98***	1.93	***2.31***	**2.57**	**2.75**
SA (BIC)	**1.39**	1.70	2.09	2.42	**1.77**	**2.26**	***2.66***	3.02
Fixed (AIC)	1.58	***1.65***	***1.82***	**1.95**	***1.81***	2.34	2.71	***2.93***
Fixed (BIC)	1.50	1.98	2.50	2.91	***1.81***	2.33	2.80	3.24
2PL	1.50	1.98	2.50	3.01	***1.81***	2.33	2.80	3.24
Semi-random	1.49	1.96	2.43	2.81	***1.81***	2.33	2.80	3.23

Note Percentages refer to the number of non-standard true item response models. *SA* Simulated annealing; fixed = best out of all $k = 0$, $k = 1$, $k = 2$, models according to AIC or BIC; 2PL = two-parameter logistic. The best RIMSE value in each column is in bold, the second best is in bold and italics

AIC had poor performance in this single cell of the design versus the 2PL and semi-random model. Finally, SA tended to do better than use of fixed k, though this trend tended to hold within a particular information criterion. For instance, SA with AIC tended to do better than fixed k with AIC selection, and SA with BIC did better than fixed k with BIC selection.

Flagging of Non-standard Items. Although a secondary outcome, we might expect that better fitting models using SA will tend to have non-standard items modeled using $k > 0$. We therefore examined *sensitivity* = # hits/# actual positives = # non-standard items using MP/# non-standard items, and the *false positive rate* = # false positives/# actual negatives = # standard items using MP/# standard items.

We desire high sensitivity, but low false positives—the upper-left region of each panel in Fig. 2. A starting temperature of 5 had a slight advantage over 10, which in turn was better than 25. A better sensitivity/false positive trade-off appears present under complete versus missing data. AIC (versus BIC) resulted in higher sensitivity, but also more false positives. It is difficult to further compare AIC and BIC due to little overlap on each axis. In some cases BIC had near zero false positives, but enough sensitivity to improve IRF recovery. For BIC and a starting temperature of 5, only two cells had false positive rates above .02 (both complete data, 80% non-standard, with .16 and .19). Excluding these two cells, sensitivity for BIC still ranged from .07 to .41. Although not explicitly depicted, a lower percentage of non-standard items tended towards the lower left of these plots, and increasing percentages are connected by straight lines. That is, a higher percentage of non-standard items tended to result in higher sensitivity and higher false positives.

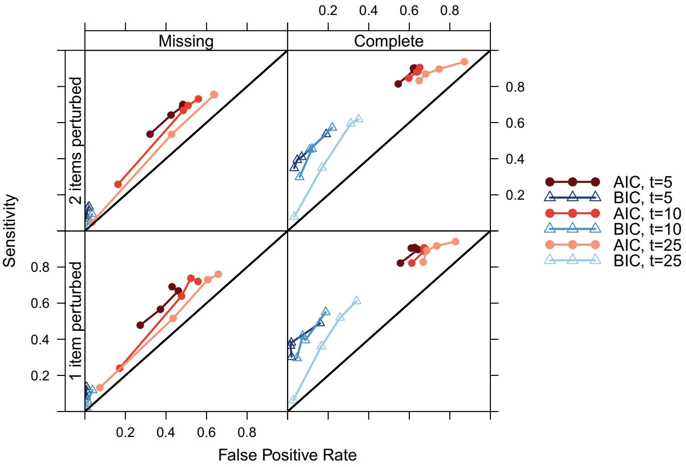


Fig. 2 False positives and sensitivity for final models selected by simulated annealing. *Note* "t" indicates starting temperature

3 Discussion and Conclusion

We conclude that SA has potential to aid in selecting polynomial order for MP-based item models in that SA tended to improve IRF recovery under most conditions. This result is promising given our initial attempt at SA implementation. For instance, tuning of the cooling schedule may further improve performance. In retrospect, a starting temperature of 25 may allow initial acceptance of many poorly fitting models, and a lower starting temperature is preferable. The number of iterations could also be increased, yet a computational advantage is still apparent over a step-wise approach: 300 fitted models would have only allowed change in polynomial order for 3 items on a test with $n = 100$.

There were some trade-offs in the choice of AIC versus BIC. AIC tended to have greater gains in IRF recovery, except under missing data and when few items followed a non-standard model. As AIC had greater sensitivity in modeling non-standard items with an MP, it also tended to result in some over-fitting. Given the great contrast in sensitivity and false positive rates, we suppose that the psychometrician's preference for a conservative (BIC) versus liberal (AIC) flagging of non-standard items may guide which to use. Of course, other optimality criterion or use of other item fit statistics may be used in future research. In addition, test length, sample size, and the amount of missing data may also be important to consider and could be further examined.

A similar computational problem may hold for other flexible parametric modeling techniques (Lee & Bolt, 2018): Should we use a standard item model or a different modeling approach? To the extent that the test is very long, this same problem may occur if one is trying to decide between several different models for *each* test item. Of course, substantive theory should be used to guide modeling choices where possible. However, in the absence of good theory, an automated approach such as that we have provided here may be a reasonable start to help improve fit while identifying which items require closer examination, especially for a long test or large item bank. MP-based models do not directly inform about the source of item misfit. Further follow-up analyses with alternative models and/or content analysis of particular items may provide insight into whether an MP or other modeling approach is appropriate. That is, there is both room for MP-based item models to complement other modeling approaches, and also for such combinatorial optimization algorithms to be used in selecting whether to use any of these other modeling approaches.

Acknowledgements We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [funding reference number RGPIN-2018-05357]. Cette recherche a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), [numéro de référence RGPIN-2018-05357].

References

Birnbaum, A. (1968). *Some latent trait models and their use in inferring an examinee's ability* (pp. 395–479). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Černý, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, *45*, 41–51.

Drezner, Z., & Marcoulides, G. A. (1999). Using simulated annealing for selection in multiple regression. *Multiple Linear Regression Viewpoints*, *25*, 1–4.

Duncan, K. A., & MacEachern, S. N. (2013). Nonparametric Bayesian modeling of item response curves with a three-parameter logistic prior mean. In M. C. Edwards & R. C. MacCallum (Eds.), *Current topics in the theory and application of latent variable models* (pp. 108–125). New York, NY: Routledge.

Edwards, M. C., Flora, D. B., & Thissen, D. (2012). Multistage computerized adaptive testing with uniform item exposure. *Applied Measurement in Education*, *25*, 118–114.

Falk, C. F. (2018). *A monotonic polynomial graded response model*. Presentation at the International Test Commission Conference, Montreal, Canada.

Falk, C. F., & Cai, L. (2016a). Maximum marginal likelihood estimation of a monotonic polynomial generalized partial credit model with applications to multiple group analysis. *Psychometrika*, *81*, 434–460.

Falk, C. F., & Cai, L. (2016b). Semi-parametric item response functions in the context of guessing. *Journal of Educational Measurement*, *53*, 229–247.

Feuerstahler, L. (2016). Exploring alternate latent trait metrics with the filtered monotonic polynomial IRT model. Ph.D. thesis, Department of Psychology, University of Minnesota.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 671–680.

Lee, S., & Bolt, D. M. (2018). Asymmetric item characteristic curves and item complexity: Insights from simulation and real data analyses. *Psychometrika*, *83*, 453–475.

Lee, Y. S. (2007). A comparison of methods for nonparametric estimation of item characteristic curves for binary items. *Applied Psychological Measurement*, *31*, 121–134.

Liang, L., & Browne, M. W. (2015). A quasi-parametric method for fitting flexible item response functions. *Journal of Educational and Behavioral Statistics*, *40*, 5–34.

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods*, *9*(3), 354–368.

Molenaar, D. (2015). Heteroscedastic latent trait models for dichotomous data. *Psychometrika*, *80*(3), 625–644.

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kickpatrick, R. M., et al. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, *81*, 535–549.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50–64.

Pritikin, J. N. (2016). rpf: Response probability functions. https://CRAN.R-project.org/package=rpf, r package version 0.53.

R Core Team. (2015). R: A language and environment for statistical computing. http://www.R-project.org, ISBN 3-900051-07-0.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*(4), 611–630.

Ramsay, J. O., & Wiberg, M. (2017). A strategy for replacing sum scoring. *Journal of Educational and Behavioral Statistics*, *42*(3), 282–307.

Stander, J., & Silverman, B. W. (1994). Temperature schedules for simulated annealing. *Statistics and Computing*, *4*, 21–32.

Woods, C. M. (2007). Empirical histograms in item response theory with ordinal data. *Educational and Psychological Measurement*, *67*, 73–87.

Xu, X., & Douglas, J. A. (2006). Computerized adaptive testing under nonparametric IRT models. *Psychometrika*, *71*(1), 121–137.

TestGardener: A Program for Optimal Scoring and Graphical Analysis

Juan Li, James O. Ramsay and Marie Wiberg

Abstract The aim of this paper is to demonstrate how to use TestGardener to analyze testing data with various item types and explain some main displays. TestGardener is a software designed to aid the development, evaluation, and use of multiple choice examinations, psychological scales, questionnaires, and similar types of data. This software implements the optimal scoring of binary and multi-option items, and uses spline smoothing to obtain item characteristics curves (ICCs) that better fit the real data. Using TestGardner does not require any programming skill or formal statistical knowledge, which will make optimal scoring and item response theory more approachable for test analysts, test developers, researchers, and general public.

Keywords Item response theory · Graphical analysis software · Optimal scoring · Spline smoothing

1 Introduction

TestGardener is the successor of TestGraf, and both softwares are designed to aid the development, evaluation, and use of multiple-choice examinations, psychological scales, questionnaires, and similar types of data. TestGraf was developed by James Ramsay (1995) and has been widely used as an analysis and/or teaching tool of nonparametric item response theory (IRT) in fields like education (Liane, 1995; Nering & Ostini, 2010), psychology (Lévesque et al., 2017; Sachs et al., 2003), medicine (Gomez, 2007; Luciano et al., 2010), and business (Laroche et al., 1999). Users who are familiar with TestGraf can still choose to use its algorithms such as item correct score and kernel smoothing within the TestGardener framework. But

J. Li (✉)
Department of Mathematics and Statistics, McGill University, Montreal, Canada
e-mail: lijuan.640@gmail.com

J. O. Ramsay
Department of Psychology, McGill University, Montreal, Canada

M. Wiberg
Department of Statistics, USBE, Umeå University, Umeå, Sweden

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_8

this paper will focus on the new features (spline smoothing and optimal scoring) and displays that are included in TestGardener.

When we analyze and evaluate real-world testing data, a known problem with parametric IRT is the inability to model all items in a test accurately, even in carefully developed large-scale tests. Using spline smoothing, TestGardener can provide estimated item characteristic curves (ICC) with high precision, even for ill-behaved items. TestGardener also implements optimal scores, which consider the interaction between the test-takers' performance and the sensitivity of the items.

With the user-friendly interface and self-explanatory displays, TestGardener is designed for users with various backgrounds, with or without knowledge in IRT, statistics, and programming. Psychometricians, researchers, test developers, and teachers can easily upload their data, and have the analysis results displayed in diagrams.

TestGardener is relatively fast when analyzing real-world testing data. A sample of 54,033 test takers response data who took the quantitative part of the Swedish Scholastic Assessment Test (SweSAT) is used to demonstrate TestGardener. The SweSAT is a multiple-choice college admissions test, with a verbal and a quantitative part, each containing 80 items. The whole analysis of this 54,033*80 multi-choice data, including reading and writing files, takes about five minutes using a laptop with intel i7 core.

The next section briefly introduces the algorithms of spline smoothing and optimal scoring, which are implemented in TestGardener. The following section provides a short demo of using this software and describes some of the main displays. This paper ends with a short discussion about different versions of TestGardener, new features that may be implemented in later version, and some closing remarks.

2 Theories Behind TestGardener

The real-world testing data rarely meets all the assumptions made in the parametric IRT model. Taking one administration of SweSAT (quantitative part) as an example, the distribution of sum scores is much more skewed than the normal distribution (Fig. 1), indicating that it was a difficult test.

Furthermore, the highlighted ICCs show the ill-behaviors of some items: some items have almost linear ICCs (see highlighted curve in Fig. 2), which means that these items are not very discriminating at any ability level. There are also items with plateaus for a certain score range (Fig. 2); it means these items have no sensitivity for test takers in these ranges. It's probably because test takers with certain level of related knowledge can rule out some of the distractors and choose among the rest options. ICCs of these ill-behaved items can be difficult to estimate using parametric IRT. But TestGardener, using spline smoothing, can estimate these 80 curves without any problems and in only a few seconds.

It is important that the test scores should estimate the test takers' ability as precisely as possible, since tests and test scores are often used to make decisions about test

Fig. 1 Distribution of sum scores of one administration of the SweSAT. The histogram indicates the number of test takers within each score range, the black smooth line indicates the smooth density function. The vertical dotted lines are the 5, 25, 50, 75, and 95% quantile lines

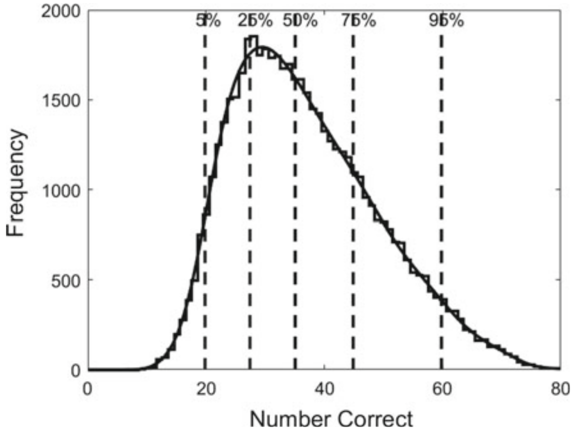
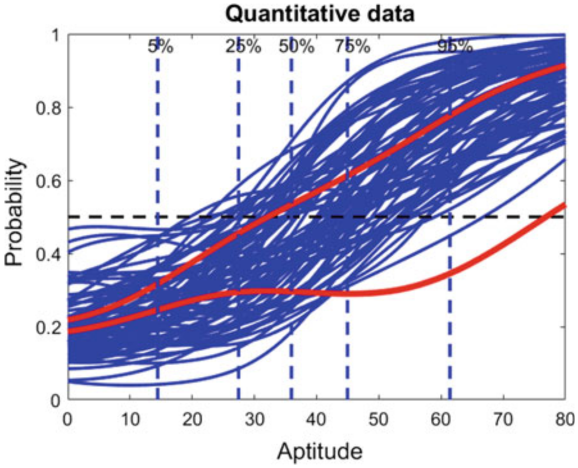


Fig. 2 Estimated ICCs of one administration of the SweSAT, quantitative part. Blue curves are the ICCs of all the 80 items, and red curves highlighted some of the ill-behaved items



takers. Sum score (or number correct score) has been the most commonly used test score because it is easy to interpret and computationally fast. However, sum scores assume that a certain item has constant sensitivity over the entire ability range, which is seldom true. For example, an easy item can have high discrimination power for lower-end test takers, but provides virtually no information about the top students, and vice versa. Optimal scoring, first proposed by Ramsay and Wiberg (2017a), considers the interaction between performance/ability-level and item sensitivity and provides more precise estimation of the test takers' ability. In their 2017 paper, Ramsay and Wiberg only considered the binary response (0/1); but with the extra information of which (wrong) option has been chosen, we can have even more precise estimation of the ability θ because sometimes some wrong options are more wrong than others.

Let S_j denote sum scores and let $P_i(\theta)$ be the probability that the test taker j with ability level θ answer an item i correctly, $i = 1, \ldots, n$; $j = 1, \ldots, N$. The estimate

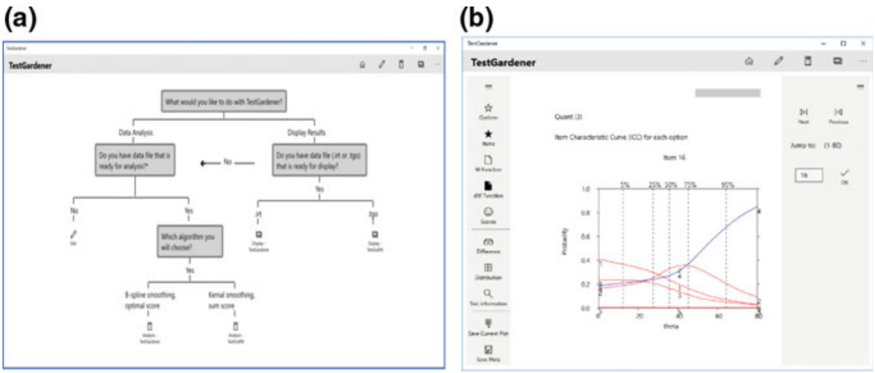

(a)
TestGardener
What would you like to do with TestGardener?
Data Analysis
Display Results
Do you have data file that is ready for analysis?*
No
Do you have data file (.irt or .tgo) that is ready for display?
No
Yes
Yes
.irt
.tgo
Which algorithm you will choose?
Yes
B-spline smoothing, optimal score
Kernal smoothing, sum score
(b)
TestGardener
Options
Items
W Function
dW Function
Scores
Difference
Distribution
Test information
Save Current Plot
Save Plots
Quant (3)
Item Characteristic Curve (ICC) for each option
Item 16
5%
25%
50%
75%
95%
1.0
0.8
0.6
0.4
0.2
0.0
Probability
0
20
40
60
80
theta
Next
Previous
Jump to: (1-80)
16
OK

Fig. 3 The opening page (**a**) and display page (**b**) of TestGardener

of optimal scores focuses on estimating the more convenient choice $W_i(\theta)$ as it will facilitate the estimations (Ramsay & Wiberg 2017a). $W_i(\theta)$ is the log-odds of $P_i(\theta)$, which can be defined in terms of $P_i(\theta)$ as

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

If U_{ij} is test taker j response to item i and if either $P_i(\theta)$ or their counterparts $W_i(\theta)$ are known or we can condition on estimates on them, then the optimal θ associated with the negative log likelihood satisfies the equation

$$\sum\nolimits_i^n \sum\nolimits_m^{M_i} [U_{ji,m} - P_{i,m}(\theta_j)] \frac{\mathrm{d}W_{i,m}}{\mathrm{d}\theta_j} = 0, \tag{2}$$

where $m = 1, \ldots, M_i$ and M_i is the number of options of item i. More details about optimal scoring can be found in Ramsay and Wiberg (2017a, 2017b) and Wiberg, Ramsay, and Li (2018). More papers about optimal scoring of multi-choice items and scale items are currently under preparation.

3 A Short Demo of TestGardener

Using TestGardener requires no knowledge of programming; users can simply upload their data (in format described in the manual) and have it analyzed. A result file in .irt format will be generated by the software; it stores all the analysis results and will be used to generate the graphical displays. Figure 3a shows the opening page of TestGardener. By following the flow chart, users should be able to find the appropriate function. Users have the option to change values of several important parameters; but for most users, they are recommended to run the analysis with default values.

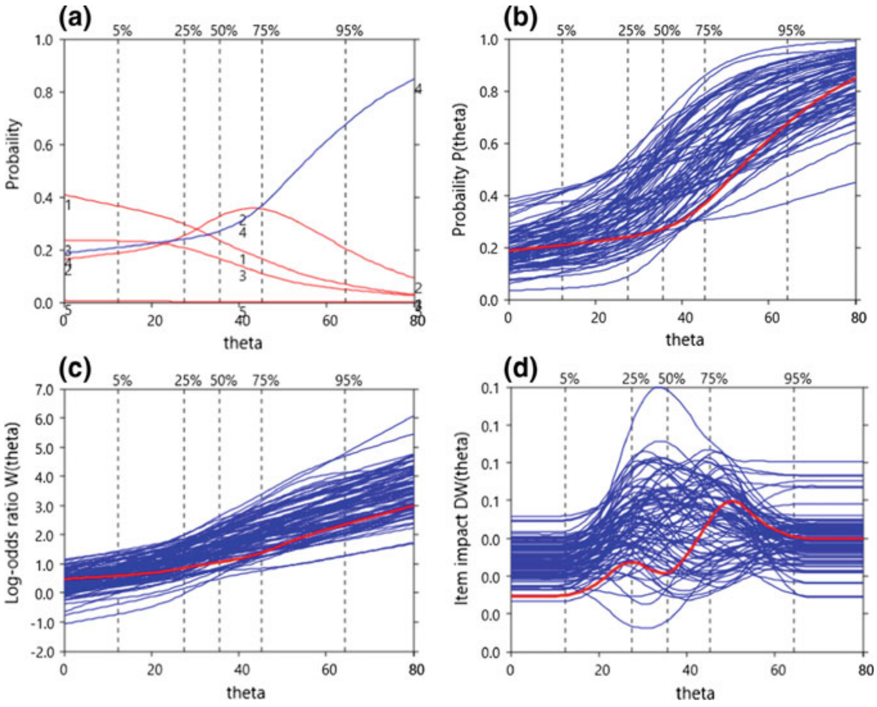


Fig. 4 ICC and other displays indicating performance of item 16. **a** ICCs of all options, blue curve indicates the right option, and red curves represent wrong options. Display **b**–**d** are for the right option only; $P_{16}(\theta)$, $W_{16}(\theta)$, and the derivative of $W_{16}(\theta)$ respectively. The blue curves are the corresponding curves of all the items, in which the curves of item 16 is highlighted with red

Using the .irt file, users can review the performance of item, test-taker, and test in various displays. The left panel in Fig. 3b lists the names of different displays, which will be introduced briefly below.

Figure 4 shows four displays that represents the performance of an individual item, here we randomly select item 16 as an example. Figure 4a shows the ICCs of all the options, where the right option is represented by a blue line. The indices associating with each curve indicate the corresponding option, so test developer or analyst can have more detailed evaluation of each option. For example, in item 16, option 2 seems quite distracting for test-takers in the middle to upper range. In fact, even for the top students, there is still around 10% probability that they may choose option 2. Figure 4b–d illustrate the probability ($P_{16}(\theta)$), the log-odds ratio ($W_{16}(\theta)$), and derivative of $W_{16}(\theta)$ ($dW_{16}(\theta)$) respectively. $W_i(\theta)$ and $dW_i(\theta)$ curves illustrate the items' sensitivity at each score value and are especially important for the process of optimal scoring. With the corresponding curves of all items (blue curves) in the background, users can have a more intuitive impression of how this item performed comparing with other items.

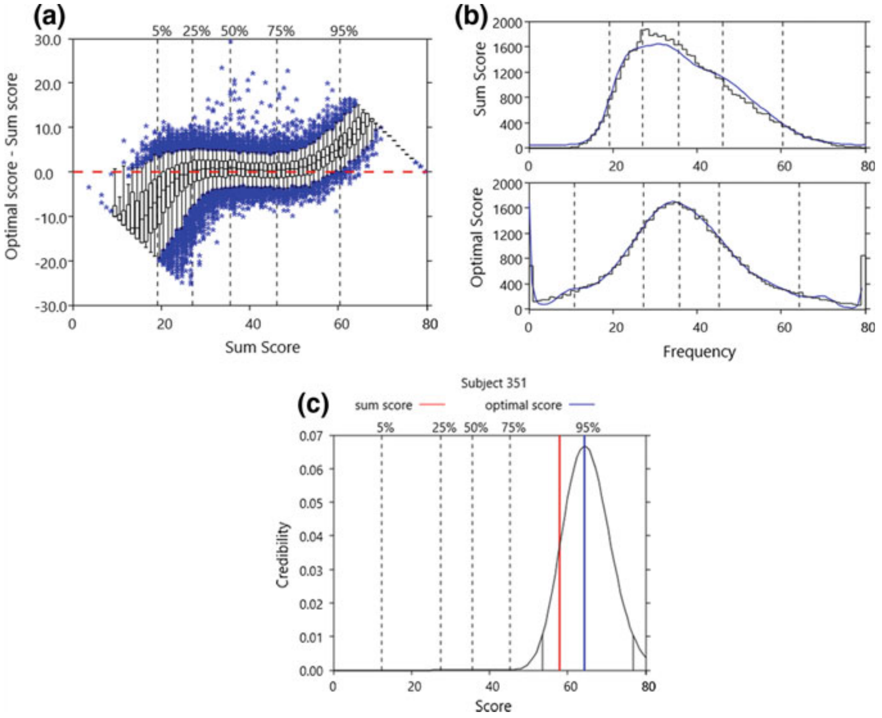


Fig. 5 Displays about the comparison between sum score and optimal score: **a** box plots of the difference between optimal score and sum score; **b** two-panel plot of the distributions of sum score and optimal score respectively; and **c** score credibility plot of subject 351: red and blue vertical lines indicate the sum score and optimal score respectively; black curve shows the likelihood (credibility) of the score, and the two black vertical lines indicate the 95% confidence interval

Figure 5 contains three displays that show the comparison between optimal score and sum score. Since the sum scores are integers, we can plot the difference between optimal score and sum score using box plots. Figure 5a shows that the differences in the middle range are distributed around zero, while for the lower and upper end, the differences are mostly negative and positive respectively. For lower end, optimal scores are corrected for guessing; while for upper end, optimal scores eliminate the influence of ill-behaved items. Figure 5b shows the distribution of sum score and optimal score, with quantile lines changing accordingly. It conveys the same information as the boxplot but from another perspective. Figure 5c shows the comparison at the individual level using the likelihood curve. Optimal score is always at the peak of the likelihood curve, hence optimal.

4 Discussion

TestGardener currently has two versions: one stand-alone version for windows system and one web-based version that can be used on any major browsers. These two versions share almost the same features, but currently, the stand-alone version has more ability to edit and prepare the data file. The web-based version is newly developed for users on other operating systems or someone who wants to try some of the features before downloading the software. It also serves as the teaching platform for optimal scoring and even item response theory, by including pages for software manual, theories, and resources. Both versions of TestGardener are still under development, although a beta version dedicating to the workshop held in Umea, Sweden this August has been published. Readers who are interested in TestGardener are welcome to try the web-based version on http://testgardener.azurewebsites.net/. But please note that this version is premature and not very stable.

The formally released TestGardener (both versions) are expected to be even faster, with more display options. For example, users may be able to choose the options that plot the confidence interval and data points on the ICCs. Plots in Fig. 4 are currently displayed separately, and we plan to implement this four-panel plot like Fig. 4 in the later version of TestGardener.

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

References

Gomez, R. (2007). Australian parent and teacher ratings of the DSM-IV ADHD symptoms differential symptom functioning and parent-teacher agreement and differences. *Journal of Attention Disorders*. *11*(1), 17–27.

Laroche, M., Chankon, K., & Tomiuk, M. (1999). Irt-based item level analysis: an additional diagnostic tool for scale purification. In J. E. Arnould, L. M. Scott (Eds.) *Advances in consumer research* (Vol 26, pp. 141–149). Provo, UT: Association for Consumer Research.

Lévesque, D., Sévigny, S., Giroux, I., & Jacques, C. (2017). Gambling-related cognition scale (GRCS): Are skills-based games at a disadvantage? *Psychology of Addictive Behaviors, 31*(6), 647–654.

Liane, P. (1995). A comparison of item parameter estimates and ICCs produced with TESTGRAF and BILOG under different test lengths and sample sizes. The University of Ottawa, thesis.

Luciano, J., Ayuso-Mateos, J., Aguado, J., Fernandez, A., Serrano-Blance, A., Roca, M., et al. (2010). The 12-item world health organization disability assessment schedule II (WHO-DAS II): A nonparametric item response analysis. *BMC Medical Research Methodology, 2010*(10), 45.

Nering, M. L., & Ostini, R. (2010). *Handbook of polytomous item response theory models*. New York: Taylor and Francis.

Ramsay, J. O. (1995). *TestGraf—a program for the graphical analysis of multiple choice test and questionnaire data [computer software]*. Montreal: McGill University.

Ramsay, J. O., & Wiberg, M. (2017a). A strategy for replacing sum scores. *Journal of Educational and Behavioral Statistics, 42*(3), 282–307.

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

Sachs, J., Law, Y., & Chan, C. K. (2003). A nonparametric item analysis of a selected item subset of the learning process questionnaire. *British Journal of Educational Psychology 73*(3), 395–423.

Wiberg, M., Ramsay, J. O., & Li, J. (2018). Optimal scores as an alternative to sum scores. In: M. Wiberg, S. Culpepper, R. Janssen, J. González, D. Molenaar (eds) *Quantitative Psychology. IMPS 2017*. Springer Proceedings in Mathematics & Statistics, vol 233. Cham: Springer.

Item Selection Algorithms in Computerized Adaptive Test Comparison Using Items Modeled with Nonparametric Isotonic Model

Mario Luzardo

Abstract A computerized adaptive test (CAT) is used in this paper where the item bank is calibrated by using the nonparametric isotonic model proposed by Luzardo and Rodríguez (Quantitative psychology research. Springer International Publishing, Switzerland, pp. 99-108, 2015). The model is based on the estimation of the inverse of the item characteristic curves (ICC), and it uses a two-stage process. First, it uses the Ramsay nonparametric estimator of the ICC (Ramsay In Psychometrika 56:611–630, 1991) and then it estimates the density function of the inverse ICC by using Ramsay's estimator. By integrating the density function and then symmetrizing it, we obtain the result. Xu and Douglas (Psychometrika 71:121–137, 2006) studied the possibility of using Ramsay's nonparametric model in a CAT. They explored the possible methods of item selection but they did not use Fisher's maximum information method because the derivatives of the ICC may not be estimated well. We present, for the isotonic model, a suitable way to estimate the derivatives of the ICCs and obtain a formula for item information that allows us to use the maximum information criterion. This work focuses on comparing three methods for selecting items in the CAT: random selection, the maximum Fisher information criterion with the isotonic model, and the Kullback-Leibler information criterion.

Keywords Isotone IRT nonparametric model · Kullback-Leibler information · Computerized adaptive test

1 Introduction

Nonparametric item response models have been an alternative to parametric item response models, especially when it comes to finding a flexible model for ICC modelling. However, a common problem is how to make CAT administration and, in particular, automatic item selection operational.

M. Luzardo (✉)
School of Psychology, University of the Republic, Montevideo, Uruguay
e-mail: mluzardo@psico.edu.uy

Eastern Regional University Center, University of Republic, Maldonado, Uruguay

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_9

Xu and Douglas (2006) explored the possibility of applying CAT by using Ramsay's nonparametric model. Under this model, the usual methods for estimating the ICC derivative do not work properly and the derivative may be negative for some values of the ability. This means it is impossible to use the maximum Fisher information criterion when choosing the items to be managed. Xu and Douglas (2006) propose as alternative the use of procedures based on Shannon entropy (Cover & Thomas, 1991; Shannon, 1948) and Kullback-Leibler information (Chang & Ying, 1996), since the implementation of these procedures does not require ICC derivatives. In addition, when the test size is large enough, they are equivalent to the maximum Fisher information criterion. The authors used a simulation study to show that both procedures work properly and have very similar outcomes.

In this paper we will show that when the nonparametric isotonic model is used to estimate the ICCs, their derivatives can be calculated in a simple way, and they can be used to estimate the Fisher information for each item. Our aim is therefore to compare this new approach with those proposed by Xu and Douglas (2006). Since the Kullback-Leibler procedure and that based on the Shannon entropy produce very similar results, we will only use Kullback-Leibler. Our main intention is to compare performances in the case of small test sizes, since the Kullback-Leibler and the Shannon entropy procedures are asymptotically equivalent to the maximum information criterion.

2 One-Dimensional Isotonic Model

The isotonic model presented in Luzardo & Rodríguez (2015) estimates the ICC in two stages. The first stage uses the Ramsay model (1991) as a preliminary estimate of the ICC, and the second obtains the isotonic estimator.

Let X be a dichotomous item and assume that $P(\theta)$ is the probability that a subject with ability θ will respond to item X correctly. As the random variable X is Bernoulli, it follows that $P(\theta) = E(X|\Theta = \theta)$, that is, the ICCs match a conditional expectation. On this basis, Ramsay estimated the ICCs by means of a nonparametric kernel regression estimator.

Let us assume that N subjects with a latent trait $\theta_1 \ldots \theta_N$ respond to n dichotomous items. Let us denote X_{ij} as the binary response of subject i to item j $(i = 1,\ldots, N\, j = 1,..,n)$. The kernel smoothing estimator of $P_j(\theta)$ is

$$\widehat{P_j(\theta)} = \frac{\sum_{i=1}^{N} X_{ij} K_h\left(\hat{\theta}_i - \theta\right)}{\sum_{i=1}^{N} K_h\left(\hat{\theta}_i - \theta\right)} \tag{1}$$

where the bandwidth h contemplates the trade-off between the variance of the estimator and the bias. Function K is a kernel and $K_h(\theta_i - \theta) = \frac{1}{h} K\left(\frac{(\theta_i - \theta)}{h}\right)$. In Eq. (1), $\hat{\theta}_i$ is the estimator of the i-th subject's ability. These estimates can be easily calculated by converting the empirical distribution of the sum of the subjects' scores to the scale determined by the distribution of the ability.

We will take—with no loss of generality—θ to have a uniform distribution in [0,1]. This assumption is justified by the non-identifiability of the scale. Let us assume that the distribution of the actual trait τ is $F(\tau)$ and let us consider a specific item with a strictly increasing ICC, which we will denote as *P(τ).*

If we change the variable $\theta = F(\tau)$, the function $P^*(\theta) = P\left(F^{-1}(\theta)\right) = P(\tau)$ is also the ICC of that item. It is clear that the distribution of θ is uniform in [0,1] and $P^*(\theta)$ is increasing.

Note that if $U_1, \ldots, U_T$ is a sample of independent random variables with a uniform distribution on the interval [0,1], then $\frac{1}{Th_d}\sum_{t=1}^{T} K_d\left(\frac{P^*(U_t)-u}{h_d}\right)$ is an estimator of the density of the random variable $P^*(U)$, where K_d is a kernel and h_d a bandwidth.

The density of $P^*(U)$ is $P^{*-1'}(u)\mathbb{I}_{[P^*(0),P^*(1)]}(u)$, where $\mathbb{I}$ is the indicator function. Then, $\frac{1}{Th_d}\int_{-\infty}^{\theta}\sum_{t=1}^{T} K_d\left(\frac{P^*(U_t)-u}{h_d}\right)du$ is a consistent estimator of P^{*-1} in θ (Dette, Neumeyer, & Pilz, 2006).

In order to apply the above property to our problem, let us consider a kernel K_r a bandwidth h_r, and a grid $\frac{1}{T}, \ldots, \frac{t}{T}, \ldots, 1$. Then, the Ramsay estimator of the ICC in each score is

$$\widehat{P^R}\left(\frac{t}{T}\right) = \frac{\sum_{i=1}^{N} K_r\left(\frac{\frac{t}{T}-\hat{\theta}_i}{h_r}\right)X_i}{\sum_{i=1}^{N} K_r\left(\frac{\frac{t}{T}-\hat{\theta}_i}{h_r}\right)} \tag{2}$$

Based on the above, the isotonic estimator of the inverse of the ICC in θ is:

$$\widehat{P^{*-1}(\theta)} = \frac{1}{Th_d}\int_{-\infty}^{\theta}\sum_{t=1}^{T} K_d\left(\frac{\widehat{P^R}\left(\frac{t}{T}\right)-u}{h_d}\right)du \tag{3}$$

The estimator $\widehat{P^*}$ is obtained by the reflection of $\widehat{P^{*-1}}$ with respect to the bisector of the first quadrant.

3 Item Selection Method Through Maximum Information

The maximum information procedure is based on the fact that when the maximum likelihood method is used to estimate the ability, the test information is inversely proportional to the estimation error of θ. It is therefore reasonable to present in the next step the item that will maximize the accumulated information. This procedure will be adopted in this article. It is therefore necessary to be able to correctly estimate the derivative of the ICC of the nonparametric isotonic model.

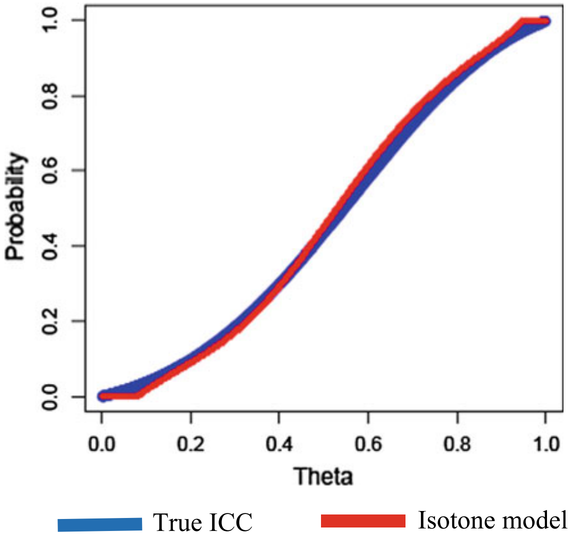


Fig. 1 True ICC and estimated ICC using the isotone model

In our case, we easily obtain a simple expression for the derivative of the ICC which is smooth and always positive by applying the inverse function derivative theorem.

$$P^{*'}(\theta) = \frac{1}{\left(P^{*-1}\right)'(P^{*}(\theta))} = \frac{Th_d}{\sum_{t=1}^{T} K_d\left(\frac{P^{*}(U_t)-P^{*}(\theta)}{h_d}\right)} \tag{4}$$

Figure 1 shows the true ICC and the estimated ICC by means of the isotonic model, and Fig. 2 shows the derivatives of this ICC.

Now, on the basis of (4), we can estimate the information function of item j through:

$$\widehat{I_j(\theta)} = \frac{\left(\frac{\partial \widehat{P_j^{*}}(\theta)}{\partial\theta}\right)^2}{\widehat{P_j^{*}}(\theta)\left(1-\widehat{P_j^{*}}(\theta)\right)} = \frac{\left[\frac{Th_d}{\sum_{t=1}^{T} K_d\left(\frac{\widehat{P^{*}\left(\frac{t}{T}\right)}-P^{*}(\theta)}{h_d}\right)}\right]^2}{\widehat{P_j^{*}}(\theta)\left(1-\widehat{P_j^{*}}(\theta)\right)} \tag{5}$$

Information estimation works very well on values where item information is maximum, having distortions when we move away from that value. Our interest is focused on a setting where information is maximum, so outside that neighborhood, we can estimate information using a linear model. Figure 3 shows the information function estimate.

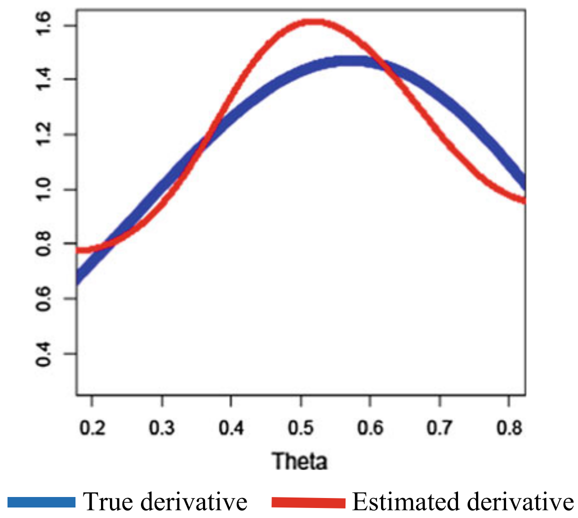


Fig. 2 True derivative and estimated derivative of the ICC

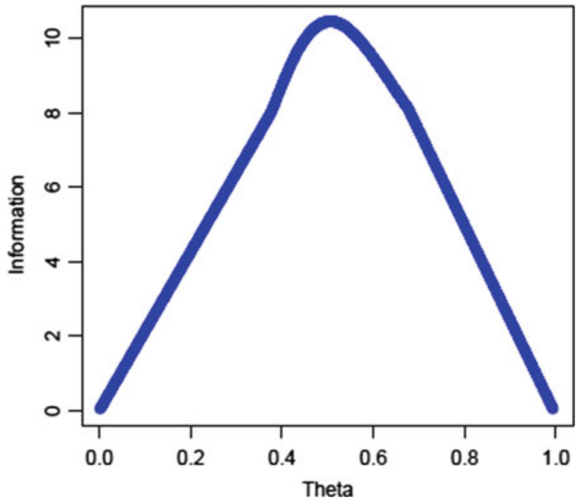


Fig. 3 Estimated information function

The maximum likelihood method will be used to estimate the ability. If $P^*(\theta)$ is the ICC when the ability follows a uniform distribution, θ will be estimated in step k through

$$\hat{\theta}_i = argmax \prod\nolimits_{j=1}^{k} P^*(\theta)^{X_{ij}} \left(1 - P^*(\theta)\right)^{1-X_{ij}} \tag{6}$$

4 Item Selection Method Using Kullback-Leibler

This divergence proposed by Kullback and Leibler (KL) (1951) measures the discrepancy between two measures of probability. On the basis of this, Chang and Ying (1996) define a measure of global information for use in CAT.

If P and Q are two probability measures over Ω, and if $\frac{dQ}{dP}$ is the Radon-Nikodym derivative of Q with respect to P, the Kullback Leibler divergence is defined as:

$$KL(P\|Q) = -\int_{\Omega} ln\frac{dQ}{dP}dP \tag{7}$$

In particular, if μ is a measure over Ω, such that f and g are densities of P and Q with respect to μ, then

$$KL(P\|Q) = \int_{\Omega} \frac{f \ \ln \ f}{g}d\mu \tag{8}$$

If we consider the maximum likelihood estimate in a parametric family *f* *(θ, x),* and $f(\theta_0, x)$ as the true density, then:

$$KL\left(f_{\theta_0}\|f_{\theta}\right) = \int f(\theta_0, x)\, ln\frac{f(\theta_0, x)}{f(\theta, x)}dx \tag{9}$$

Chang and Ying (1996) define the Kullback Leibler information for item j and subject i as

$$KL_j(\theta\|\theta_i) = E\left[ln\frac{L_j(\theta_i|X_{ij})}{L_j(\theta|X_{ij})}\right] = P_j(\theta_i)ln\frac{P_j(\theta_i)}{P_j(\theta)} + \left(1 - P_j(\theta_i)\right)ln\frac{1 - P_j(\theta_i)}{1 - P_j(\theta)} \tag{10}$$

In the context of CATs, if $\hat{\theta}_k$ is the maximum likelihood estimator of θ, after k items have been responded to, then the global information index $\mathrm{G}KL_j\left(\hat{\theta}_k\right)$ is obtained by taking the average of the discrepancy $KL_j\left(\theta\|\hat{\theta}_k\right)$ in the interval centered on $\hat{\theta}_k$, that is, if $\epsilon_k > 0$,

$$GKL_j\left(\hat{\theta}_k\right) = \int_{\hat{\theta}_k-\epsilon_k}^{\hat{\theta}_k+\epsilon_k} K_j\left(\theta\|\hat{\theta}_k\right)d\theta \tag{11}$$

The sequence $\epsilon_k \to 0$ with k. Chang and Ying (1996) recommend $\epsilon_k \propto k^{-\frac{1}{2}}$ so that the interval $\left(\hat{\theta}_k - \epsilon_k, \hat{\theta}_k + \epsilon_k\right)$ will contain the actual value of the ability. Based on the above, the item to be chosen for step (k + 1) will be the one with the greatest GKL, which has not been applied yet.

5 Simulation Study

The objective of this study was to compare three ways of selecting items in the CAT. The selection methods implemented are: the Kullback-Leibler procedure, the information-based procedure using the isotonic estimation, and random selection of items. The ability was estimated by using maximum likelihood and considering the nonparametrically estimated ICCs. Additionally, the ability for random item selection was estimated by maximum likelihood, when the ICC is estimated parametrically.

A bank of 700 items was built whose ICCs followed the two-parameter logistic model (2PL). The discrimination parameters of the items were simulated from a uniform distribution [0.75, 2.5] and the difficulty parameters were simulated from a uniform distribution [−2, 2].

To estimate the ICCs, the responses of 5000 subjects were simulated. The abilities were assumed to follow a standard normal distribution. On the basis of the responses, we used the isotonic estimator with Gaussian kernels K_r and K_d.

The bandwidths used were $h_r = (5000)^{\left(-\frac{1}{5}\right)} = 0.18$, and a robust estimate for $h_d = 0.9(5000)^{\left(-\frac{1}{5}\right)}\min\left(sd, \frac{Q_3-Q_1}{1.364}\right)$, where the deviation and the quartiles refer to the Ramsay's estimator of the ICC for each item.

For the CAT, 5000 subjects were generated whose traits had a uniform distribution on the interval [0,1]. A test of 50 items in length was applied for each of the methods and the procedures for each subtest of 5, 10, 20, 30, 40 and 50 items in length were assessed. The different procedures were compared to root mean squared error (RMSE) and bias across the simulations. The RSME and bias were computed for each subtest through:

$$RMSE = \left(\frac{\sum_{i=1}^{N}\left(\hat{\theta}_\iota - \theta_i\right)^2}{N}\right)^{\left(\frac{1}{2}\right)} \tag{12}$$

$$BIAS = \frac{\sum_{i=1}^{N}\left(\hat{\theta}_\iota - \theta_i\right)^2}{N} \tag{13}$$

Also, the RSME and bias of the estimators were calculated with a certain value of θ through:

$$RMSE(\theta) = \left(\frac{\sum_{i\epsilon I(\theta)}\left(\hat{\theta}_\iota - \theta_i\right)^2}{\#I(\theta)}\right)^{\left(\frac{1}{2}\right)} \tag{14}$$

$$BIAS(\theta) = \frac{\sum_{i\epsilon I(\theta)}\left(\hat{\theta}_\iota - \theta_i\right)^2}{\#I(\theta)} \tag{15}$$

where $I(\theta) = \{i : \theta_i = \theta,\ 1 \le i \le N\}$

6 Results

Table 1 presents the average root mean square error for the different methods and for different subtest lengths. This table shows how the procedures based on the information estimated from the isotonic ICC and the Kullback-Leibler method work in a similar way. In addition, these methods are better than the random selection of items, and the estimation of the ability is based on the isotonic nonparametric model when fewer than 30 items are administered. They are also always better than random selection and ability estimation using 2PL model. Table 2 shows a similar behavior for the bias.

Figure 4 graphically shows how the RMSE stabilizes after a test length of 20 items for the nonparametric isotone model and Kullback-Leibler procedures. The Fig. 5 shows that the bias is also stabilized.

Table 1 Average root mean square error

Selection rule	Number of items					
	5	10	20	30	40	50
Random isotone	0.203	0.147	0.108	0.085	0.074	0.074
K-L	0.154	0.125	0.098	0.085	0.076	0.074
Isotone Information	0.154	0.123	0.098	0.084	0.080	0.074
Random 2PL	0.242	0.181	0.142	0.116	0.105	0.103

Calculated on the 5000 subjects

Fig. 4 RMSE of selection procedures

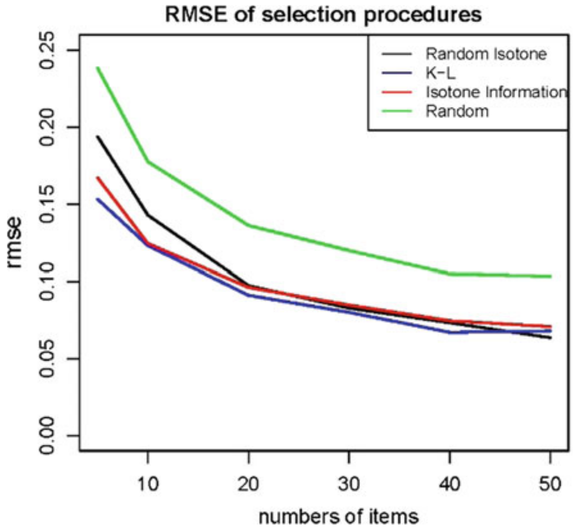


Fig. 5 Bias of selection procedures

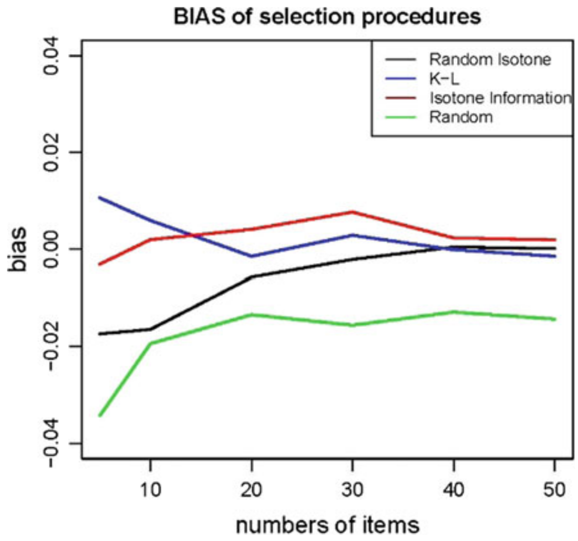
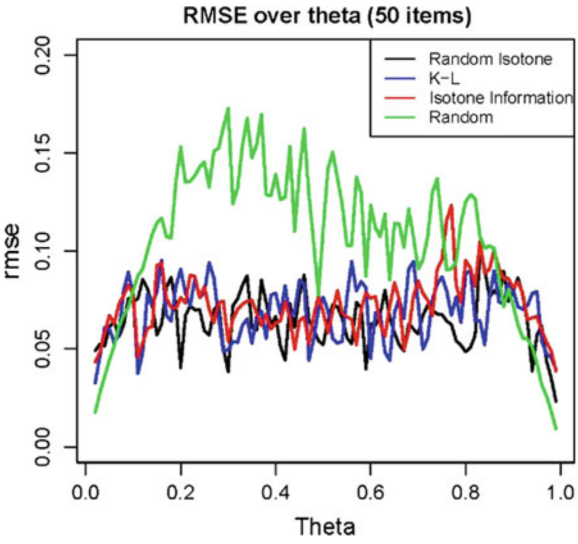


Fig. 6 RMSE over theta



An analysis of the RMSE(θ) finds that the same behavior is obtained for all θ, with an equivalence of the Kullback-Leibler method and the nonparametric isotonic method and the latter's superiority over random selection using de 2PL model procedure (Fig. 6).

When the bias is analyzed globally and as a function of θ, both methods behave appropriately (Fig. 7).

Fig. 7 BIAS over theta

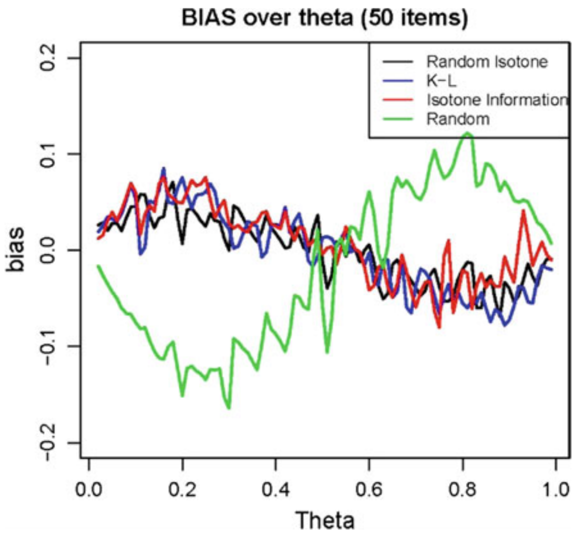


Table 2 Bias

Selection rule	Number of items					
	5	10	20	30	40	50
Random isotone	0.012	−0.013	0.003	−0.002	0.001	0.001
K-L	−0.003	−0.009	0.002	0.004	0.0001	0.001
Isotone Information	0.0004	0.003	0.007	0.005	0.005	0.005
Random 2PL	0.002	−0.014	−0.011	−0.012	−0.01	−0.01

Calculated on the 5000 subjects

7 Discussion

The procedure based on estimating information through the isotonic model quickly converges to the actual trait, stabilizing after 20 items. The performance of the procedure presented based on the isotonic model is similar to that of KL in terms of root mean square error, and simpler to implement. It is also observed that both adaptive procedures work better than random selection of items in terms of root mean square error.

It would be wise to expand some studies that would extend these results, for example by studying the rate of exposure of the items, which it has been omitted in this work.

References

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

Cover, H. H., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.

Dette, H., Neumeyer, N., & Pilz, K. (2006). A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli, 12*(3), 469–490.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694.

Luzardo, M., & Rodríguez, P. (2015). A nonparametric estimator of a monotone item characteristic curve. In *Quantitative Psychology Research* (pp. 99–108). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-19977-1_8.

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

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal, 27,* 379–423.

Xu, X., & Douglas, J. (2006). Computerized adaptive testing under nonparametric IRT models. *Psychometrika, 71,* 121–137.

Utilizing Response Time in On-the-Fly Multistage Adaptive Testing

Yang Du, Anqi Li and Hua-Hua Chang

Abstract On-the-fly multistage adaptive testing (OMST), which integrates computerized adaptive testing (CAT) and multistage testing (MST), has recently gained popularity. While CAT selects each item on-the-fly and MST bundles items to pre-assembled modules, OMST assembles modules on-the-fly after the first stage. Since item selection algorithms play a crucial role in latent trait estimation and test security in CAT designs, given the availability of response time (RT) in the current testing era, researchers have been actively striving to incorporate RT into item selection algorithms. However, most such algorithms were only applied to CAT whereas little is known about RT's role in the domain of OMST. Building upon previous research on RT-oriented item selection procedures, this research intends to apply RT-oriented item selection algorithms to OMST. This study found that the relative performance of RT-oriented item selection methods in OMST was consistent with CAT. But the underlying item bank structure and test design features can make a huge difference with respect to estimation accuracy and test security.

Keywords On-the-fly multistage tests · Response time · CAT · Test security · Item bank usage

Y. Du (✉)
University of Illinois, Urbana-Champaign, 1310 S 6th St, Education Building Rm. 188R, Champaign, IL 61820, USA
e-mail: yangd2@illinois.edu

A. Li
University of Illinois, Urbana-Champaign, 603 East Daniel St., 418 Psych Bldg, Champaign, IL 61820, USA
e-mail: anqili4@illinois.edu

H.-H. Chang
Purdue University, Steven C. Beering Hall of Liberal Arts and Education,100 N. University Street, West Lafayette, IN 47907, USA
e-mail: chang606@purdue.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_10

1 Introduction

Before response time (RT) became available to test developers and researchers, item responses were the sole information source of test-takers and test items. Within the scope of item responses, traditional item selection methods in computerized adaptive tests (CAT) have been focused on improving estimation accuracy, test security, as well as item bank usage.

Given the rapid development of test technology, RT attracted more attention of psychometricians. Owing to the indispensable support of RT models, such as the log-normal model, and the hierarchical framework of response and RT (van der Linden 2006; van der Linden, Breithaupt, Chuah, & Zhang, 2007), RT has been effectively utilized in a myriad of aspects, such as detecting cheating, rapid guessing, or student disengagement, e.g., Wang and Xu (2015). In other words, apart from item responses, researchers are interested in understanding how RT can provide additional information with respect to the latent traits of test-takers.

Due to the exponential growth of RT research, the purpose of item selection algorithms in CAT has also been expanded. In addition to the aforementioned orientations, researchers, such as Fan, Wang, Chang, and Douglas (2012) and Choe, Kern, and Chang (2018), have incorporated RT to a number of item selection methods to shorten the test time span while maintaining estimation accuracy and test security. However, both of these studies only explored RT's role in conventional CAT, in which items are sequentially selected and administered. Little is known about RT's role in multistage tests, especially in on-the-fly multistage tests (OMST). This study is aimed at investigating the role of RT-oriented item selection algorithms in OMST. To be more specific, this study intends to answer the following questions: How do RT-oriented methods perform in OMST? Does item bank stratification impact RT-oriented item selection in OMST?

The following parts of this paper will first introduce the theoretical framework of CAT, OMST, and RT models. Next, a description of item selection algorithms and simulation studies will be presented. The results and conclusions will be shown at the end.

2 Adaptive Testing Design

CAT is developed on the basis of various measurement models in the item response theory (IRT) framework, among which the 3PL model is universally used Lord and Novick (1968). In the 3PL model, the estimates of the examinees' latent traits are usually obtained by maximum likelihood estimate(MLE). However, MLE has two main drawbacks. On the one hand, at the beginning of the test, when only a handful of items are administered, MLE is not stable and accordingly the corresponding latent traits estimates may not be accurate. On the other hand, when examinees' all responses are correct or incorrect, their MLEs would be practically meaningless.

To compensate for these weaknesses, an alternative estimator, expected a posteriori (EAP), is therefore employed Bock and Mislevy (1982).

In conventional CAT selection procedures, in order to minimize the variance of latent trait estimates, maximizing Fisher information (MI) is commonly used Lord and Novick (1968). However, this method is notorious for improperly prioritizing more discriminating (high a-parameter) items Chang, Qian, and Ying (2001), Chang and Ying (1999) and Hau and Chang (2001). Despite such items' efficiency in estimating latent traits, their skewed exposure rates, consequently, compromise test security. To balance the item pool usage and thus to improve test security, Chang et al. (2001) proposed a-stratification with b-blocking (ASB) procedure. To be specific, according to the magnitude of b-parameters, the item bank is first divided into several blocks. Next, within each block, items are stratified into several strata based on the magnitude of a-parameters. Finally, items within the same strata are grouped across the blocks. In item selection, low-a items are selected at the beginning whereas high-a items are selected at a later stage of the test.

As a special case of CAT, multistage testing (MST) has gained its long-standing popularity in the testing industry Yan, von Davier, and Lewis (2014). Unlike CAT which estimates examinees' abilities item by item, MST sequentially routes examinees to pre-assembled modules of items Yan et al. (2014). Nevertheless, since the modules and panels in MST are pre-assembled, it is not cost-effective to recruit many employees to review these pre-assembled modules Zheng and Chang (2015). To further improve MST, Zheng and Chang (2015) proposed OMST which assembles modules on-the-fly after the first stage. In OMST, item selection methods in MST are still applicable. The only difference is that within each stage of OMST, a set of items are sequentially selected on-the-fly to optimize the objective selection algorithms. Therefore, items are no longer bundled together. The estimation efficiency and test security can be significantly enhanced.

3 Response Time Model

As RT data can be collected conveniently via test delivery software, it has drawn greater attention with its potential as a rich source of information. Among a variety of RT models, the lognormal model within the hierarchical modeling framework by van der Linden et al. (2007) has received the most recognition for its practicality. In this model, it is assumed that each examinee operates at a constant latent speed unless the exam is over-speeded van der Linden et al. (2007). Specifically, the lognormal model characterizes the RT of a fixed person on a set of test items van der Linden (2006). The model assumes that the RT of the ith person on item j is the realization of a random variable t_{ij} following a lognormal distribution, which can be written as:

$$f(t_{ij}|\tau_i) = \frac{\alpha_j}{t_{ij}\sqrt{2\pi}} e^{-[\alpha_j(\log t_{ij} - \beta_j + \tau_i)]^2/2}, \tag{1}$$

where τ_i is the latent speed of the ith person, α_j and β_j respectively denote time discrimination and time intensity. β_j and τ_i are fixed on the same scale. Accordingly, the MLE of latent speed τ_i is given by

$$\hat{\tau}_i = \frac{\sum_{j=1}^{k} \alpha_j^2(\beta_j - \log t_{ij})}{\sum_{j=1}^{k} \alpha_j^2}. \tag{2}$$

In addition, the expected RT based upon latent speed is given as

$$E(T_{ij}|\tau_i) = e^{\beta_j - \tau_i + 1/2\alpha_j^2}. \tag{3}$$

4 Item Selection Methods

In this study, three RT methods proposed by Fan et al. (2012) and Choe et al. (2018) were employed. They are a-stratification b-blocking with time (ASBT), maximum information with beta matching (MIB), and generalized maximum information with time (GMIT). To evaluate the performance of these methods, ASB and MI were adopted as two baseline methods.

The ASB method, shown below, prioritizes items which minimize the absolute difference between item difficulty parameter and the examinee's estimated latent trait in a stratified item bank.

$$B_j(\hat{\theta}_i) = \frac{1}{|\hat{\theta}_i - b_j|}. \tag{4}$$

MI selects the next item that maximizes the Fisher information given as

$$I_j(\theta_i) = a_j^2(\frac{1 - P_j(\theta_i)}{P_j(\theta_i)})(\frac{P_j(\theta_i) - c_j}{1 - c_j})^2 \tag{5}$$

By incorporating RT in item selection criterion, ASBT tends to minimize the expected RT while matching the examinee's latent trait with the item difficulty parameter Fan et al. (2012).

$$BT_j(\hat{\theta}_i, \hat{\tau}_i) = \frac{B_j(\hat{\theta}_i)}{E(T_{ij}|\hat{\tau}_i)} = \frac{\frac{1}{|\hat{\theta}_i - b_j|}}{e^{\beta_j - \tau_i + 1/2\alpha_j^2}} \tag{6}$$

MIB proposed by Choe et al. (2018) favors items with greater information and a shorter distance between latent speed and item time intensity.

$$IB_j(\hat{\theta}_i, \hat{\tau}_i) = \frac{I_j(\hat{\theta}_i)}{|\beta_j - \hat{\tau}_i|} \tag{7}$$

GMIT is based upon maximum information with time (MIT) proposed by Fan et al. (2012), which tends to maximize the information while minimizing the expected RT. Choe et al. (2018) added two variables trying to adjust its performance, which is given by

$$IT_j^G(\theta_i, \tau_i) = \frac{I_j(\theta_i)}{|E(T_{ij}|\tau_i) - v|^w} \tag{8}$$

Since locating the optimal v and w is not the primary goal of this study, only $v = 0$ and $w = 1$ were adopted.

5 Simulation Design and Evaluation Criteria

In this study, all simulations were completed in R and 50 replications were conducted in total. The simulated OMST includes three stages, within each of which fifteen items were administered. Moreover, in the first stage, due to the concern of test overlap rates, 20 parallel modules were pre-assembled based on the target information criteria Luecht (2000). The difficulty levels of all these parallel modules were anchored at zero, where items produced the largest Fisher information. 1000 examinees and an item bank of 500 items were simulated. The parameters of examinees and items are shown below.

- Item Bank

$$(a_j^*, b_j, \beta_j) \sim \mathcal{MVN}[\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1}], \qquad \boldsymbol{\mu_1} = \begin{bmatrix} 0.3 \\ 0.0 \\ 0.0 \end{bmatrix}, \qquad \boldsymbol{\Sigma_1} = \begin{bmatrix} 0.10 & 0.15 & 0.00 \\ 0.15 & 1.00 & 0.25 \\ 0.00 & 0.25 & 0.25 \end{bmatrix},$$

$$c_j \sim \beta\ [2, 10], \qquad \alpha_j \sim U\ \ [2, 4],$$

  where $a_j^* = log\ a_j$. In other words, a_j follows a lognormal distribution.
- Examinees

$$(\theta_i, \tau_i) \sim \mathcal{MVN}[\boldsymbol{\mu_2}, \boldsymbol{\Sigma_2}], \qquad \boldsymbol{\mu_2} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \boldsymbol{\Sigma_2} = \begin{bmatrix} 1.00 & 0.25 \\ 0.25 & 0.25 \end{bmatrix}.$$

To investigate the relationship between RT-oriented item selection methods and item bank stratification, our simulation studies were administered in both stratified and unstratified item banks. The ASB item bank structure was employed Chang et al. (2001). Due to our three-stage test structure, our item bank was accordingly divided into three disjoint sub-banks, which served as sub-pools for each stage, respectively. Each sub-pool had 167, 167, and 166 items, respectively. A summary of item selection methods can be found in Table 1.

The performance of item selection procedures was assessed in four aspects: estimation accuracy, test efficiency and stability, test security, as well as item bank usage efficiency. For each of these aspects, the following measures were adopted respectively:

Table 1 A summary of item selection methods

Item bank	Selection methods	Selection criterion
Stratified item bank	ASB	$B_j(\hat{\theta}_i) = \frac{1}{\lvert\hat{\theta}_i - b_j\rvert}$
	ASBT	$BT_j(\hat{\theta}_i, \hat{\tau}_i) = \frac{B_j(\hat{\theta}_i)}{E(T_{ij}\lvert\hat{\tau}_i)}$
	MIB	$IB_j(\hat{\theta}_i, \hat{\tau}_i) = \frac{I_j(\hat{\theta}_i)}{\lvert\beta_j - \hat{\tau}_i\rvert}$
	GMIT	$IT_j^G(\theta_i, \tau_i) = \frac{I_j(\theta_i)}{\lvert E(T_{ij}\lvert\tau_i) - v\rvert^w}$
Unstratified item bank	MI	$I_j(\theta_i) = a_j^2(\frac{1 - P_j(\theta_i)}{P_j(\theta_i)})(\frac{P_j(\theta_i) - c_j}{1 - c_j})^2$
	MIB	$IB_j(\hat{\theta}_i, \hat{\tau}_i) = \frac{I_j(\hat{\theta}_i)}{\lvert\beta_j - \hat{\tau}_i\rvert}$
	GMIT	$IT_j^G(\theta_i, \tau_i) = \frac{I_j(\theta_i)}{\lvert E(T_{ij}\lvert\tau_i) - v\rvert^w}$

- Estimation accuracy measures: root mean squared error (RMSE) of examinees' latent traits and latent speeds.

$$RMSE(\hat{\theta}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2} \tag{9}$$

$$RMSE(\hat{\tau}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\tau}_i - \tau_i)^2} \tag{10}$$

- Test efficiency and stability measures: mean and standard deviation of test time tt_i.

$$\bar{tt} = \frac{1}{n}\sum_{i=1}^{n} tt_i = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} t_{ij} \tag{11}$$

$$s_{tt} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(tt_i - \bar{tt})^2} \tag{12}$$

- Test security measures: the exposure rates of individual items er_j as well as mean and standard deviation of test overlap rates $tor_{ii'}$ (Chen et al., 2003).

$$er_j = \frac{k}{n}, \tag{13}$$

$$\bar{tor} = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{i'=i+1}^{n} tor_{ii'} = \frac{n}{L(n-1)} \sum_{j=1}^{m} er_j^2 - \frac{1}{n-1}, \tag{14}$$

$$s_{tor} = \sqrt{[\binom{n}{2} - 1]^{-1} \sum_{i=1}^{n-1} \sum_{i'=i+1}^{n} (tor_{ii'} - \bar{tor})^2}, \tag{15}$$

where k is the times of item j being administered, n is the number of examinees. m is the item bank size and L is the fixed test length. Note that the simulated OMST has a fixed length of 45 items, thus the mean of the exposure rates will not be informative. For the simplicity of our evaluation, this study chose 0.2 as the maximum proper exposure rate, a commonly set value for high-stakes tests. Items whose exposure rates are greater than 0.5 were treated as over-exposed.

– Item Bank Usage Efficiency: raw counts of unused items and over-exposed items.

6 Results

Stratified Item Bank. The results of the stratified item bank are shown in Tables 2 and 3. It is apparent that ASB, ASBT, and MIB performed similarly in estimating examinees' latent traits and latent speeds. In comparison, GMIT produced better estimation accuracy with strikingly shorter test time. The other two RT item selection procedures also had more efficient tests compared to the baseline ASB method. However, in contrast to the other three methods, GMIT almost doubled their overlap rates, suggesting that every two students may share one half of identical test items. The overlap rates of MIB and ASBT were similar to each other, and their test security was acceptable. In terms of the item bank usage, GMIT resulted in over 315 unused items and 26 over-exposed items, greatly impairing the item bank usage efficiency and test security. The item bank usage of ASBT and MIB were, again, similar to each other. Although their item bank usage was more efficient than GMIT, there were still quite a few items staying in the bank either never selected or over-exposed.

Unstratified Item Bank. The results for the unstratified item bank are shown in Tables 4 and 5. It was expected that MI would produce the most accurate estimation. GMIT's estimation accuracy was comparable to MI. In addition, consistent with our

Table 2 Estimation accuracy, test efficiency, and overlap rates in the stratified item bank

	RMSE ($\hat{\theta}$)	RMSE ($\hat{\tau}$)	tt	tt.sd	tor	tor.sd
ASB	0.287	0.049	59.822	29.981	0.248	0.219
ASBT	0.291	0.049	52.464	26.337	0.266	0.226
MIB	0.296	0.049	55.283	17.247	0.252	0.214
GMIT	0.267	0.050	44.466	23.086	0.434	0.349

Table 3 Exposure rates and item bank usage in the stratified item bank

	Unused items	er < 0.5	0 < er < 0.2
ASB	126	11	341
ASBT	126	11	326
MIB	126	11	342
GMIT	315	26	97

Table 4 Estimation accuracy, test efficiency, and overlap rates in the unstratified item bank

	RMSE ($\hat{\theta}$)	RMSE ($\hat{\tau}$)	tt	tt.sd	tor	tor.sd
MI	0.238	0.049	55.310	27.517	0.283	0.289
MIB	0.302	0.0489	54.722	16.002	0.147	0.161
GMIT	0.246	0.0485	43.887	22.616	0.339	0.312

Table 5 Exposure rates and item bank usage in the unstratified item bank

	Unused items	er > 0.5	0 < er < 0.2
MI	220	6	187
MIB	0	0	455
GMIT	254	17	161

results for the stratified item bank, GMIT and MIB both improved the test stability by reducing the mean and standard deviation of test time. As for the test security, GMIT still produced relatively higher overlap rates, more unused items, and more over-exposed items.

In summary, to answer the first research question, how RT-oriented item selection methods perform in OMST, the results are shown below:

– Estimation accuracy: in terms of latent trait estimation, GMIT outperformed the other RT item selection methods in both item banks. The estimation accuracy of latent speeds of all RT methods were comparable.
– Test efficiency and stability: all RT-oriented methods in both item banks had better test efficiency and stability, with shorter average test time and smaller test time standard deviation. This is reasonable as the other two baseline methods, ASB and MI, only selected items within the scope of item accuracy. Among RT item selection methods, in accordance with the results in Choe et al. (2018), MIB performed similarly to ASBT. GMIT, again, produced the most efficient test.
– Test security: the estimation accuracy and test security trade-off mentioned in Choe et al. (2018) was manifest in this study. The security of GMIT-based tests, implied by their overlap rates and exposure rates, was far from satisfactory whereas MIB performed similarly to ASBT and their corresponding test security was maintained well in both item banks.

- Item bank usage efficiency: among RT methods, GMIT had the worst item bank usage efficiency. MIB and ASBT had more efficient item bank usage.

To answer the second research question, i.e., if item bank stratification impacts item selection in OMST, the results of MIB and GMIT were plotted in Fig. 1. Based on the plot, it is apparent that item bank stratification influenced RT item selection in OMST: in contrast to the stratified item bank, both MIB and GMIT had more accurate estimation, lower overlap rates as well as better item bank usage in the unstratified item bank. Such results may seem to be contradictory to what Chang et al. (2001) advocated; Yet note that item bank stratification is primarily designed to balance item exposure. But it may not successfully control the overlap rates or improve the item bank usage efficiency. Moreover, when the item bank is stratified and the item selection algorithms take more parameters into computing, the stratified sub-banks will undoubtedly have fewer items satisfying the algorithms. This would result in more over-exposed items and higher overlap rates.

Overall, in OMST, all RT item selection methods improved the test efficiency by shortening the test time. The trade-off between estimation accuracy and test security was evident: while GMIT produced the most accurate estimation compared to ASBT and MIB, its test security was also severely undermined. Additionally, in OMST, the stratified item bank structure may also tend to hinder the test security and item bank usage.

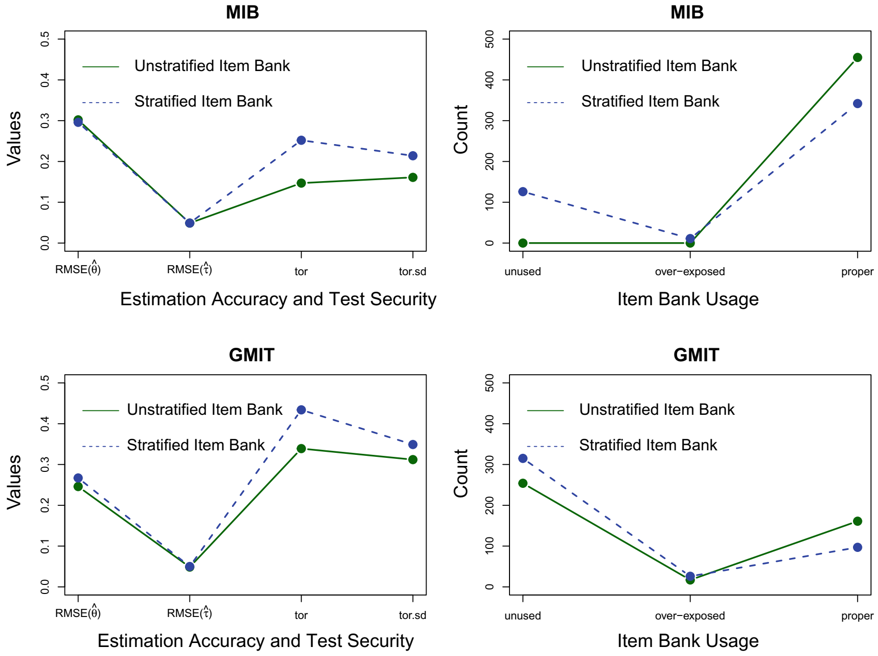


Fig. 1 Impacts of item bank stratification on RT methods in OMST

7 Conclusions

This study adopted RT-oriented item selection algorithms proposed by Fan et al. (2012) and Choe et al. (2018). We tested their performance in OMST within both stratified and unstratified item banks. The performance of ASBT, MIB, and GMIT in OMST were consistent with what Choe et al. (2018) found in terms of estimation accuracy and test stability. Yet the item bank stratification influenced the test security and item bank usage efficiency in OMST.

This study is limited in that only a handful of conditions of examinees, item bank, item parameters, and test designs were included. Such limitations may make our conclusions less generalizable. In future studies, more conditions of item and person parameters should be examined.

Acknowledgements The authors thank Dr. Dylan Molenaar for his detailed and helpful comments. Send request for reprints or further information to Yang Du at yangd2@illinois.edu or Anqi Li at anqili4@illinois.edu.

References

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431–444.

Chang, H.-H., Qian, J., & Ying, Z. (2001). A-stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement*, *25*(4), 333–341.

Chang, H.-H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 211–222.

Chen, S.-Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement*, *40*(2), 129–145.

Choe, E. M., Kern, J. L., & Chang, H.-H. (2018). Optimizing the use of response times for item selection in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *43*, 135–158.

Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. (2012). Utilizing response time distributions for item selection in cat. *Journal of Educational and Behavioral Statistics*, *37*(5), 655–670.

Hau, K.-T., & Chang, H.-H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement*, *38*(3), 249–266.

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

Luecht, R. M. (2000). *Implementing the computer-adaptive sequential testing (cast) framework to mass produce high quality computer-adaptive and mastery tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181–204.

van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, *44*(2), 117–130.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477.

Yan, D., von Davier, A. A., & Lewis, C. (2014). *Computerized multistage testing: Theory and applications*. Taylor & Fransis Group.

Zheng, Y., & Chang, H.-H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, *39*(2), 104–118.

Heuristic Assembly of a Classification Multistage Test with Testlets

Zhuoran Wang, Ying Li and Werner Wothke

Abstract In addition to the advantages of shortening test and balancing item bank usage, multistage testing (MST) has its unique merit of incorporating testlets. Testlet refers to a group of items sharing the same piece of stimulus. As MST can include an entire testlet in one module, fewer stimuli are required than items. On the other hand, computerized adaptive testing (CAT) selects item one by one, thus excludes the possibility of several items sharing the same stimulus. In this way, testlets in MST save the stimuli processing time and facilitate ability estimate. In order to utilize the advantages brings by testlet, a classification MST was designed to upgrade an operational listening test. A heuristic module top-down assembly procedure incorporating testlet was developed based on the modified normalized weighted absolute deviation heuristic (NWADH). A three-stage classification MST with 1-3-5 panel design was assembled to classify examinees into six levels. A real data-based simulation study was conducted to compare the performance of the classification MST and the operational linear test in terms of ability recovery and classification accuracy. The bi-factor model was used in item parameter calibration and examinee scoring. Results show the 30-item MST had a similar performance as the 44-item linear test with prior knowledge of examinee ability and outperformed the 44-item linear test without prior information, in both ability recovery and classification accuracy. In conclusion, the classification MST can shorten the test while keeping a good accuracy.

Keywords Multistage testing · Classification · Testlets

Z. Wang (✉)
University of Minnesota Twin Cities, Minneapolis, MN, USA
e-mail: wang5105@umn.edu

Y. Li
American Councils for International Education, Washington, DC, USA

W. Wothke
Werner Wothke Consulting, Washington, DC, USA

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_11

1 Introduction

Multistage testing (MST) is increasingly used in the language testing area (Yan, von Davier, & Lewis, 2016; Breithaupt, Ariel, & Veldkamp, 2005). In addition to the advantages of shortening test and balancing item bank usage, MST has its unique merit of incorporating testlets. Testlet refers to a group of items sharing the same piece of stimulus. For instance, when the stimulus is a longer segment of spoken language, the items could reflect two or three unrelated questions about that segment. The entire package of stimulus and associated items then form the testlet. As an MST can include an entire testlet in one module, fewer stimuli are required than items. On the other hand, another type of adaptive testing, computerized adaptive testing (CAT), selects items one by one, thereby excluding the possibility of several items sharing the same stimulus (Boyd, Dodd, & Fitzpatrick, 2013). In this way, testlets in MST save the stimuli processing time and facilitate ability estimation (Zheng, Chang, & Chang, 2013).

Like all the other testing formats, MST can also be used to classify test takers. A common example would be a certification test, in which test takers are either classified as pass or fail, or arranged into one of several levels. In this study, a classification MST was constructed based on the administration of a linear language test to explore the performance of MST in practice. As there was no algorithm to cope with testlets in classification MST, a module assembly algorithm incorporating testlets was needed. The heuristic module top-down assembly was proved to work well in MST without testlets (Zheng, Nozawa, Zhu, & Gao, 2015). Thus, in this study a heuristic module assembly procedure taking testlets into account was created and used to develop the MST.

In addition to generating MST, this assembly algorithm can also be applied in linear test generation. Due to concerns over test security, some large-scale admission tests, such as the Scholastic Aptitude Test (SAT), still utilize a linear test format. When testlets are involved, the linear test assembly is no longer simple. Among the other issues, testlet structure should be taken into consideration. The assembly algorithm developed in this study can be used to choose testlets when the number of items and the number of testlets are set beforehand.

Please note, in MST studies sometimes “testlet” and “module” are used interchangeably. In this study, testlet represents the set of items sharing the same stimuli, while module stands for a group of items to be shown in one stage at one time.

Table 1 Proficiency levels of items

Proficiency level	0+	1	1+	2	2+	3
Number of items	41	68	36	28	14	11

2 Methods

2.1 Bi-factor Testlet Model

A bi-factor testlet model was used to calibrate item parameters as well as to conduct the simulation study (Bradlow, Wainer, & Wang, 1999). The model can be written as

$$P_{ij}(\theta_i) = \frac{1}{1 + \exp\left(-1.7\left(a_0\theta_{i0} + a_k\theta_{ik} - b_j\right)\right)} \tag{1}$$

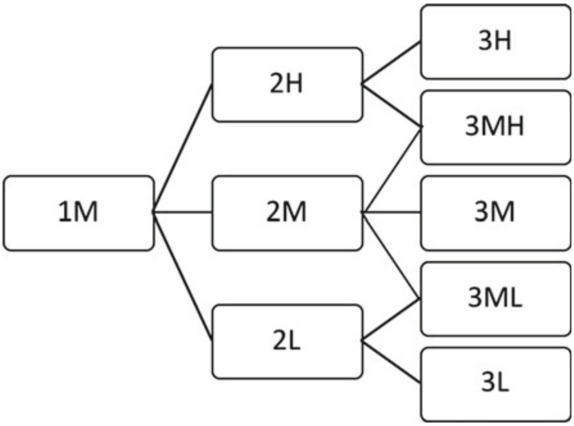
where, $P_{ij}(\theta_i)$ is the probability test taker i correctly answers item j. a_o and a_k are the discrimination parameters of the primary dimension and testlet k, respectively. Due to sample size restrictions, the constraint of equal primary dimension discrimination parameters was imposed. Thus, all items share the same a_o. In the testlet model, all the discrimination parameters (loadings) on the same testlet (specific dimension) are restricted to be equal. Thus, each testlet k only has one discrimination parameter a_k. θ_{i0} and θ_{ik} are the primary ability and the testlet k-related trait of test taker i, respectively, while b_j is the difficulty of item j.

2.2 Data Cleaning

The linear test aims at classifying test takers into one of the six Interagency Language Roundtable (ILR) proficiency levels. There are 15 test forms, either classifying test takers among the four easy levels (0+, 1, 1+, 2), or classifying test takers among the four hard levels (1+, 2, 2+, 3). Whether a test taker receives an easy form or a hard form is based on the teacher's rough evaluation. Each form contains 44 items. Altogether, 2247 test takers answered 337 items. As item response theory (IRT) was used to estimate item parameters, only items with more than 100 responses were retained to construct the MST. There were 198 items with more than 100 responses. 2144 test takers answered these items. The item proficiency distribution is shown in Table 1.

Among the 198 items, there are 34 two-item testlets and four three-item testlets. The remaining 118 items which do not belong to any testlets were designated single items.

Fig. 1 MST panel



2.3 MST Panel Design

In order to classify test takers into six levels, the final stage should contain five modules (Zheng, Nozawa, Zhu, & Gao, 2016). Thus, a 1-3-5 panel design was used as shown in Fig. 1.

In the 30-item MST, each stage contained 10 items. To ensure the item bank usage balance, there was no parallel form for stage-3 modules. There were two and five parallel forms, respectively, for the stage-2 and stage-1 modules. No overlap was allowed between modules. 160 out of 198 items were utilized to build the MST. Backward assembly was used, so the modules in stage-3 were assembled first. The anchors for the five modules in stage-3 were set at the five classification cutoff points, which were the 66-th percentiles of *b* values in level “1”, “1+”, “2”, “2+” and “3” items, respectively. The anchor for each module in the preceding stage was set to the average of the anchor values of all the modules in the next stage accessible from this module (e.g. The anchor value of 2H is the mean of anchor values of 3H and 3 MH.). The routing cut score was set to be the intersection of module information curves in the adjacent two follow-up modules (e.g. the routing cut score for 3MH and 3M from 2 M is the intersection of module information curves in 3 MH and 3M).

2.4 Heuristic Module Assembly with Testlets

Panels are assembled from modules based on the panel design. As the testlet constraint is at the module level, the bottom-up assembly can be utilized to construct panels from modules, with which the modules are freely mixed and matched to create panels (Breithaupt, et al., 2005). The other panel assembly method is the bottom-up assembly, which is used when there are panel level restrictions.

Heuristic assembly was employed to construct modules from item banks, as it has the advantage of simplicity and feasibility compared to linear programming (van der Linden, 1998). The new algorithm incorporating testlets was based on the modified normalized weighted absolute deviation heuristic (NWADH; Zheng, et al., 2016; Luecht, 1998). There are two constraints: the module information value at the anchor, and the anchor itself. The ten items in each module constitute eight testlets/single items, either seven single items plus one three-item testlet, or six single items plus two two-item testlets. Let $i = 1, 2, \ldots, I$ denote one of the I testlets in the item bank, $j = 1, 2, \ldots, J$ denote one of the J testlets needed in a module. The jth testlet in the module was selected by maximizing:

$$\sum_{i=1}^{I} \sum_{n=1}^{N} e_{i,n} x_i \tag{2}$$

subject to $x_i \in \{0, 1\}$ and $\sum_{i=1}^{I} x_i = J$. Here x_i is the decision variable for selecting testlet i into the module. Then $e_{i,n}$, the "priority index" of testlet i on constraint n, is defined as:

$$e_{i,n} = \begin{cases} 1 - \frac{d_{i,n}}{\sum_{i \in R_{j-i}} d_{i,n}}, \; if \sum\limits_{i \in R_{j-i}} d_{i,n} \neq 0; \\ 0, \qquad\qquad otherwise. \end{cases} \tag{3}$$

where, R_{j-1} is the subset of item bank excluding the $j - 1$ already selected testlets, and $d_{i,n}$ is an "absolute deviation".

For the constraint on module information at the anchor(s), $n = 1$ and

$$d_{i,1} = \left| \frac{T_1 - \sum_{k=1}^{I} u_{k,1} x_k}{L - l + 1} - u_{i,1} \right|, \; i \in R_{j-1} \tag{4}$$

where, T_1 is the target value of module information, and $u_{i,1}$ is the average item information of testlet i at the anchor location. Based on Weiss and Gibbons' strategy (Weiss & Gibbons, 2007), only the information in the primary dimension was considered here. L is the number of items needed to be selected into a module; l is the module length after testlet i is selected.

For the constraint on item difficulty, $n = 2$ and

$$d_{i,2} = \left| T_2 - u_{i,2} \right|, \; i \in R_{j-1} \tag{5}$$

where, T_2 is the target item difficulty, and $u_{i,2}$ is the average difficulty of testlet i.

2.5 Simulation Design

500 test takers were simulated from a $N(0, 1)$ distribution. To facilitate comparison, each test taker took 5 tests: a 30-item MST, a 44-item linear test with prior information, a 30-item linear test with prior information, a 44-item randomly selected test, and a 30-item test randomly selected test. As MST is intended to shorten the test, 44-item MST was not considered. Test taker responses were simulated using the bi-factor testlet response model.

In MST, the bi-factor testlet response model was used to estimate examinee ability at the end of each stage. After stage-1 and stage-2, the examinees were distributed to the next stage based on their ability estimate in the primary dimension (Weiss & Gibbons, 2007).

A 44-item and a 30-item linear test both with prior information about students' language levels were used to imitate the operational test. A hard version test and an easy version test were assembled from the same item bank. The hard test contained items from levels "1+", "2", "2+", and "3", while the easy test contained items from levels "0+", "1", "1+", and "2". The same number of items (for the 30-item test, 7 items were selected from each of the two extreme levels, while 8 items were selected from each of the two central levels) were randomly selected from each of the four levels considering the testlet structure. An examinee's prior classification function was used to simulate the assignment of tests by teachers. The probability of been assigned to take the hard test is

$$P_i(\theta_{0i}) = \frac{1}{1+\exp(d\theta_{0i})} \tag{6}$$

where, θ_{0i} is the ability of examinee i in the primary dimension, d is a constant to control the discrimination ability of teachers' prior information on examinee ability. d was arbitrarily set to be 0.7 in this study.

When there is no prior information about students' language level, items are randomly selected to construct linear tests. In this way a 44-item and a 30-item linear test were used as baseline, with items randomly selected while considering testlets.

All five kinds of test (MST, two linear randomly selected tests, and two linear with prior information tests) tests were generated for each examinee, respectively. 50 repetitions were administered to ensure a stable result. Note that the linear tests used IRT item parameters, thus were different from the operational scoring procedure using sum score.

2.6 Evaluation Criteria

Bias and RMSE were used to evaluate θ_0 estimate accuracy, as it is the primary ability the tests measured. The formula of bias and RMSE are as follows

Table 2 Weight matrix

0	1	2	3
1	0	1	2
2	1	0	1
3	2	1	0

$$bias = \frac{\sum_{i=1}^{N} \hat{\theta}_{i0} - \theta_{i0}}{N} \tag{7}$$

$$RMSE = \sqrt[2]{\frac{\sum_{i=1}^{N} \left(\hat{\theta}_{i0} - \theta_{i0}\right)^2}{N}} \tag{8}$$

where, $\widehat{\theta}_{i0}$ and θ_{i0} are the estimated and true primary ability of examinee i respectively. N is the sample size.

The weighted Kappa coefficient which was used to measure classification accuracy is as follows

$$\kappa = 1 - \frac{\sum_{l=1}^{K} \sum_{m=1}^{K} w_{lm} o_{lm}}{\sum_{l=1}^{K} \sum_{m=1}^{K} w_{lm} e_{lm}} \tag{9}$$

where, K is the number of classification levels. In this study, w_{lm}, o_{lm}, and e_{lm} are elements in the weight, observed, and expected matrices, respectively. All three matrices are of size $K \times K$. The observed and expected matrices show the agreement between test-based and true classification and the hypothetical chance agreement. The diagonal elements of the weight matrix are zero, while the off-diagonal elements demonstrate the seriousness of disagreement. The elements one off the diagonal are weighted have weight 1, elements two off the diagonal are weighted have weight 2, etc. The weight matrix is shown in Table 2.

3 Results

3.1 Module Information

Figures 2, 3 and 4 depict the module information and anchor for each module.

The modules in stage-3 fitted the module design blueprint well. The peaks of different modules are divergent and close to the respective anchors. On the other hand, due to the small item bank, some modules in stage-2 and stage-1 were off-target and some others had lower module information.

3.2 Accuracy Evaluation

Table 3 shows the estimation accuracy under the five test designs. The 30-item MST has the best ability estimate accuracy. Thus, MST can decrease the test length while maintaining accurate ability estimation and classification.

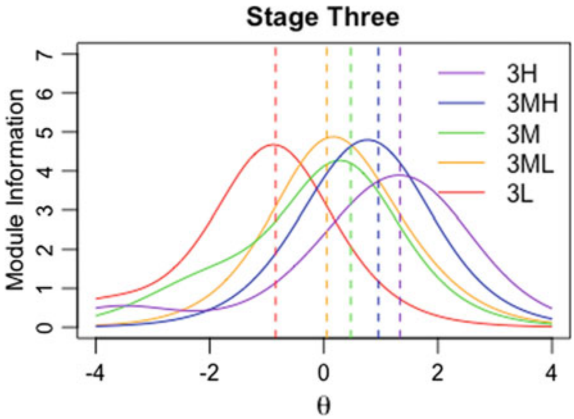


Fig. 2 Module information and anchors in stage-3

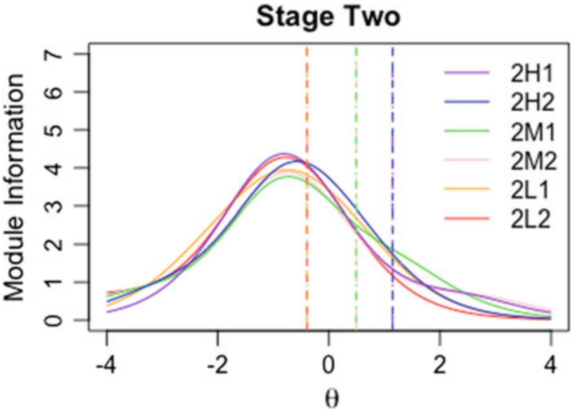


Fig. 3 Module information and anchors in stage-2

Table 3 Estimation accuracy

	Bias	RMSE	Kappa
CTT_prior_30	−0.01	0.41	0.9
CTT_prior_44	−0.01	0.32	0.92
CTT_random_30	−0.05	0.5	0.89
CTT_random_44	−0.03	0.38	0.91
MST	−0.00	0.32	0.92

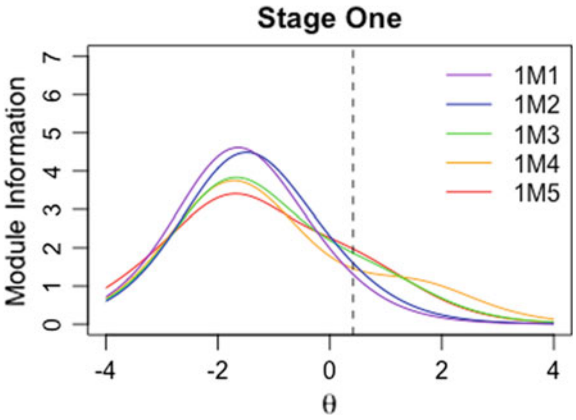


Fig. 4 Module information and the anchor in stage-1

4 Discussion

Several reasons may account for the considerably small differences among tests. First, due to the restriction of the item bank, the modules in stage-1 and stage-2 did not provide enough information at the anchors. In this way, MST design cannot bring its priority to full play. For the purpose of module assembly algorithm evaluation, in future studies larger item banks can be simulated. The advantage of MST is expected to be more obvious with the larger item bank. Second, there were only around 10 items in level 2+ and level 3, thus the high ability test takers used roughly the same sets of items in the linear test with prior information and MST. Unsurprisingly, their ability recovery and classification accuracy in the two tests were very similar. Third, with the 2PL-based testlet model, the 30-item test is practically long enough to accurately recover ability. Thus, the increment from 30 items to 44 items did not result in substantial difference in ability recovery. In a forthcoming study, shorter test length can be attempted.

References

Boyd, A. M., Dodd, B., & Fitzpatrick, S. (2013). A comparison of exposure control procedures in CAT systems based on different measurement models for testlets. *Applied Measurement in Education, 26*(2), 113–135.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*(2), 153–168.

Breithaupt, K., Ariel, A., & Veldkamp, B. (2005). Automated simultaneous assembly for multistage testing. *International Journal of Testing, 5*(3), 319–330.

Luecht, R. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement, 22*(3), 224–236.

Van der Linden, W. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement, 22*(3), 195–211.

Weiss, D. J., & Gibbons, R. D. (2007). Computerized adaptive testing with the bi-factor model. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC conference on computerized adaptive testing.*

Yan, D., von Davier, A., & Lewis, C. (Eds.). (2016). *Computerized multistage testing: Theory and applications*. CRC Press.

Zheng, Y., Chang, C. H., & Chang, H. H. (2013). Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement. *Quality of Life Research, 22*(3), 491–499.

Zheng, Y., Nozawa, Y., Zhu, R., & Gao, X. (2016). Automated top-down heuristic assembly of a classification multistage test. *International Journal of Quantitative Research in Education, 3*(4), 242–265.

Statistical Considerations for Subscore Reporting in Multistage Testing

Yanming Jiang

Abstract This study examines factors that influence the reliability of subscores and the accuracy of subscore estimates in multistage testing (MST). The factors considered in the study include the number of subtests, subtest length, correlations among subscores, item pool characteristics such as item pool size relative to the number of items required for an MST and statistical properties of items. Results indicated that the factors that most influenced subscore reliability and subscore estimates were subtest length, item pool size, and the degree of item discrimination.

Keywords Subscore · Subscore reliability · Multistage testing

1 Introduction

In K-12 large-scale state assessments, subscores are of interest to score users because of their potential diagnostic value in determining examinees' strengths and weaknesses in different subcontent areas. Such diagnostic information may be used to plan future remedial instruction.

The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) states that "scores should not be reported for individuals without appropriate evidence to support the interpretations for intended uses (Standard 6.12, pp. 119)." This applies to subscores as well. The Standard 1.14 also clarifies that when more than one score is reported for a test, the subscores that comprise the assessment should be sufficiently distinct and reliable, and be consistent with the construct(s) being assessed.

Subscore reliability is an important requirement for subscore reporting. In this study, we examine subscore reliability in multistage testing (MST). Characteristics of MSTs differ from those of linear tests in that an MST is generally shorter, with item difficulties that are adaptive to examinee abilities at various stages. This study

Y. Jiang (✉)
Educational Testing Service, Princeton, NJ 08541, USA
e-mail: yxjiang@ets.org

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_12

examines the combined impact of the characteristics of MST, item pool properties, subtest lengths, and subscore correlations on the subscore reliability and the accuracy of subscore estimates. There is currently limited research available on subscore reliabilities in an MST setting.

2 Methods

2.1 *MST Design and Multidimensional Item Response Theory (MIRT) Models*

This study focused on two-stage MST with one routing module in Stage 1 and three modules in Stage 2 with three levels of item difficulty: easy, moderate, and difficult. We assume that the total number of items is the same across subscores and that within a subscore, the total number of items is also the same in the two stages. For example, for a 2 × 16 design (two subscores and 16 items per subscore), each student takes 8 items in each subtest in Stage 1 and the same number in Stage 2.

An item pool with only dichotomously-scored items was considered. The compensatory multidimensional 2PL model with simple structures was assumed (i.e., each item belongs to one and only one dimension). Assuming there are K dimensions, the probability that individual j with ability $\theta_j = \left(\theta_{j1}, \ldots, \theta_{jK}\right)$ answers item i correctly is

$$P_i\left(\boldsymbol{\theta}_j\right) = P\left(x_{ij} = 1|\boldsymbol{\theta}_j, \boldsymbol{a}_i, \boldsymbol{b}_i\right) = \frac{1}{1 + \exp\left[-Da_{ik}\left(\theta_{jk} - b_{ik}\right)\right]}, \tag{1}$$

where x_{ij} is individual j's response value to item i and item i belongs to dimension k; a_{ik} and b_{ik} are the discrimination and difficulty parameters, respectively, for item i on dimension k; the constant D equals 1.7; $i = 1, 2,\ldots, I, j = 1, 2,\ldots, J$, and $k = 1, 2,\ldots, K$.

The latent vector, $\boldsymbol{\theta}_j$, is of dimension K in which each element is intended to measure a proficiency of a specific skill. It is assumed that each element has unit variance and correlations among dimensions are generally equal.

The expected a posteriori (EAP) estimators are used to estimate subscores. The estimation is conducted using Multidimensional Discrete Latent Trait Models (*mdltm*) software (von Davier, 2016).

The reliability of subscore k is estimated as:

$$\hat{\rho}_k = 1 - \frac{\frac{1}{J}\sum_{j=1}^{J} \sigma^2\left(\hat{\theta}_{jk}|\boldsymbol{X}_{jk}\right)}{\sigma^2\left(\hat{\theta}_k\right)}, \tag{2}$$

where $\hat{\theta}_k$ is a random variable representing a theta estimate on subtest k, $\sigma^2(\hat{\theta}_k)$ is the variance of $\hat{\theta}_k$, $\sigma^2(\hat{\theta}_{jk}|\boldsymbol{X}_{jk})$ is the estimated variance of $\hat{\theta}_{jk}$ conditional on examinee j's response pattern $\boldsymbol{X}_{jk}$, and J is the total number of examinees.

2.2 Simulation Conditions

Sinharay (2010) indicated that for a subscore to have potential for added value, a subscore has to consist of at least 20 items and be sufficiently distinct from other subscores. For an MST, we might relax this a little because of the adaptive nature of the test. Therefore, subtest lengths of both 16 and 20 are considered. The simulation study was based on a random sample of 5000 examinees from a multivariate normal distribution with specified covariance structures. For each simulation condition, there were 100 replications of item pools, the process of item selection, and item calibration and ability estimation. Detailed simulation conditions are listed in Table 1.

For each simulated item pool, automated test assembly (ATA) of an MST was conducted using *lpSolve* software in R (Diao & van der Linden, 2011; Konis, 2009), which is based on the maximin principle, that is, the minimum values for the target test information function (TIF) at target theta values are maximized. No overlapping items are allowed among Stage 2 modules. After calibration of Stage 1 items, routing an examinee into the easy module in Stage 2, for example, is based on K thresholds where the kth threshold is met by $\hat{\theta}_{jk}$ for all $k = 1, 2, \ldots, K$. These thresholds are predetermined based on the multivariate normal distributions such that the percentage of examinees who are routed into the easy module ranged approximately from 22 to 25%. The same is true for examinees routed to the difficult module in Stage 2.

Table 1 Simulation conditions

Simulation parameters	Levels
Number of subtests	2 or 3
Subtest length	16 or 20 score points
Correlation among subscores	0.5, 0.7, or 0.9
Characteristics of item pool in each subcontent area	
Percentage of items of the pool required for subtest assembly	Medium (70% usage) or large (50% usage)
Average item discrimination	0.7, 0.8, 0.9, or 1.0
Item difficulty	N(0, 0.7^2), N(0, 1), or N(0, 1.3^2)

2.3 Simulation Procedure

Below is the simulation procedure for one replication:

1. Simulate an item pool that meets specified requirements of item pool size, number of subtests, and item parameter distributions; run *lpSolve* software to assemble an MST.
2. Simulate item responses for all examinees and all items selected for the MST; calibrate items that are in the routing module using *mdltm* software and obtain initial dimensional theta estimates.
3. Route examinees into one of the three Stage 2 modules based on the initial theta estimates and a set of predetermined theta thresholds; suppress items not taken by an examinee so that the final item response matrix includes only item responses for those items that an examinee actually took.
4. Calibrate the final item response matrix and obtain final item parameter and subscore estimates.

2.4 Evaluation of Results

Subscore reliability estimates were evaluated based on their mean and standard error over 100 replications. The accuracy of subscore estimates was evaluated based on the average bias and root mean square error (RMSE) over all examinees by $\frac{1}{J}\sum_{j=1}^{J}\left[\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_{jkr}-\theta_{jk}\right)\right]$ and $\frac{1}{J}\sum_{j=1}^{J}\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_{jkr}-\theta_{jk}\right)^2}$, respectively, where θ_{jk} is the latent ability of individual j at dimension k, $\hat{\theta}_{jkr}$ is the estimate of θ_{jk} for the rth replication, and R is the total number of replications.

3 Results

Table 2, Figs. 1 and 2 are based on equal simulation conditions across dimensions and the results for all dimensions were quite similar. Therefore, results are only shown for Dimension 1 or Subscore 1.

The results in Table 2 demonstrate that long subtests, a discriminating item pool, and low subscore correlations had a positive impact on subscore reliability and subscore estimates. Compared with the 3 × 16 design, the 2 × 16 design performed slightly better.

Figure 1 contrasts the results of the amount of variations in item difficulty (in terms of the standard deviation of item difficulty parameters being 1.3 versus 0.70) for the 2 × 20 design. Large variations negatively affected subscore reliability and subscore estimates. However, the effect was small in magnitude.

Table 2 Impact of number of subtests, subtest length, average item discrimination, and subscore correlations on subscore reliability and subscore estimates

Test design	Item pool characteristics			Subscore correlation	Subscore reliability		Subscore estimates	
	Pool size	Mean a-parm[1]	b-parm[1]		Mean	SE[2]	Bias	RMSE
2 × 16	Medium	0.7	N(0, 1)	0.5	0.726	0.034	0.0124	0.4445
	(70% usage)			0.7	0.713	0.038	−0.0050	0.4511
				0.9	0.702	0.039	0.0179	0.4605
		0.8		0.5	0.772	0.022	0.0124	0.4151
				0.7	0.762	0.024	−0.0050	0.4202
				0.9	0.752	0.026	0.0179	0.4305
		0.9		0.5	0.795	0.025	0.0124	0.3981
				0.7	0.785	0.029	−0.0050	0.4036
				0.9	0.777	0.030	0.0179	0.4123
		1.0		0.5	0.823	0.015	0.0124	0.3759
				0.7	0.815	0.016	−0.0050	0.3800
				0.9	0.808	0.018	0.0179	0.3898
2 × 20		0.7		0.5	0.788	0.025	0.0124	0.4048
				0.7	0.778	0.028	−0.0050	0.4087
				0.9	0.769	0.031	0.0180	0.4193
		0.8		0.5	0.820	0.020	0.0124	0.3785
				0.7	0.811	0.022	−0.0050	0.3849
				0.9	0.803	0.024	0.0179	0.3938
		0.9		0.5	0.843	0.016	0.0124	0.3563
				0.7	0.838	0.018	−0.0050	0.3608
				0.9	0.830	0.020	0.0179	0.3701
		1.0		0.5	0.860	0.011	0.0124	0.3409
				0.7	0.854	0.013	−0.0050	0.3444
				0.9	0.849	0.014	0.0179	0.3534
3 × 16		0.7		0.5	0.723	0.042	0.0124	0.4508
				0.7	0.708	0.047	−0.0050	0.4593
				0.9	0.700	0.048	0.0179	0.4689
		0.8		0.5	0.766	0.029	0.0124	0.4227
				0.7	0.753	0.033	−0.0050	0.4315

(continued)

Table 2 (continued)

Test design	Item pool characteristics			Subscore correlation	Subscore reliability		Subscore estimates	
	Pool size	Mean a-parm[1]	b-parm[1]		Mean	SE[2]	Bias	RMSE
				0.9	0.743	0.035	0.0179	0.4432
		0.9		0.5	0.793	0.022	0.0124	0.4032
				0.7	0.784	0.024	−0.0050	0.4097
				0.9	0.774	0.026	0.0179	0.4238
		1.0		0.5	0.818	0.017	0.0124	0.3827
				0.7	0.809	0.019	−0.0050	0.3906
				0.9	0.801	0.021	0.0179	0.4025

[1] Item discrimination and difficulty parameters
[2] Standard error of subscore reliability estimates

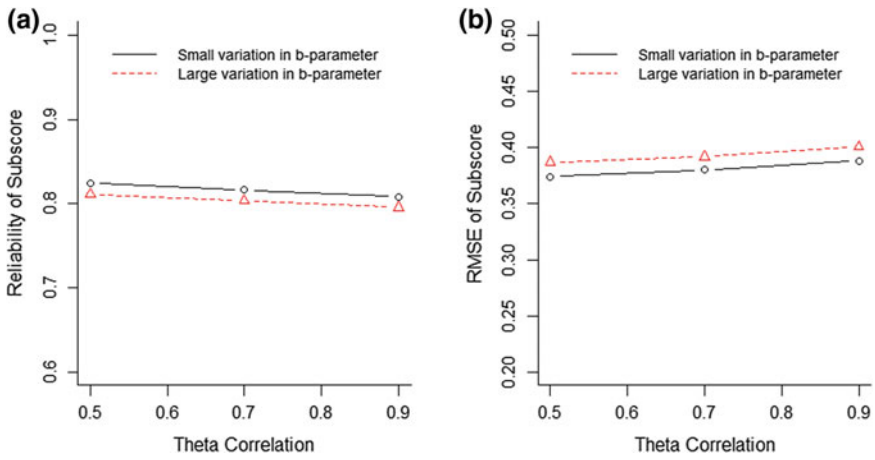


Fig. 1 Impact of variation in item difficulty (small vs. large) and subscore correlations on subscore reliability (**a**) and subscore estimates (**b**) based on 2 × 20 design

An increase in item pool size had a positive impact on both subscore reliability and subscore estimates. But the effects were larger for a less discriminating item pool than a more discriminating item pool (Fig. 2).

Table 3 presents two scenarios based on the 3 × 16 design with both equal (0.8) and unequal (0.8, 0.9, and 1.0) mean item discriminations across three dimensions and a specific theta correlation matrix.[1] Medium sized item pool and standard normal distribution for item difficulty parameter were assumed. The results indicated

[1] Theta correlation matrix is $\begin{pmatrix} 1 & 0.5 & 0.7 \\ 0.5 & 1 & 0.9 \\ 0.7 & 0.9 & 1 \end{pmatrix}$.

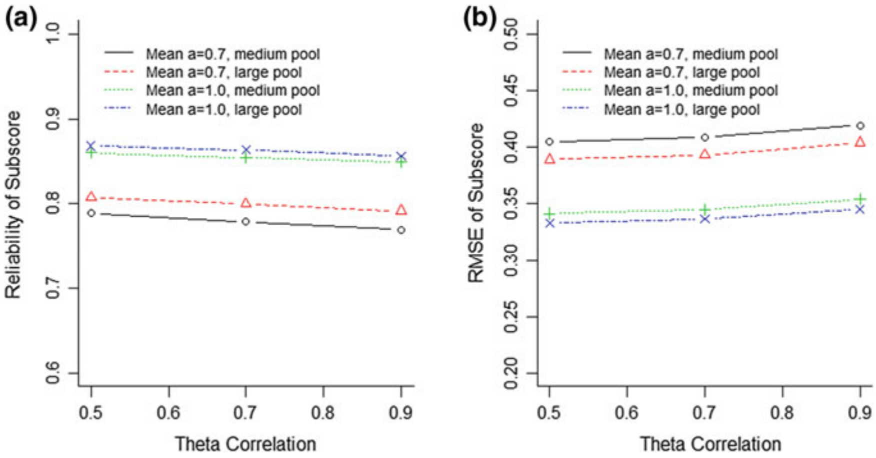


Fig. 2 Impact of item pool size (medium vs. large) and mean item discriminations (0.7 vs. 1.0) on subscore reliability (**a**) and subscore estimates (**b**) based on 2 × 20 design

Table 3 Subscore reliability and subscore estimates under both equal and unequal mean item discriminations across dimensions based on 3 × 16 design

Mean item discrimination	Subscore	Subscore reliability		Subscore estimates	
		Mean	SE[1]	Bias	RMSE
0.8	1	0.760	0.031	0.012	0.427
	2	0.752	0.030	0.012	0.435
	3	0.740	0.035	0.015	0.439
Mixed[2]	1	0.759	0.028	0.012	0.428
	2	0.779	0.024	0.012	0.412
	3	0.802	0.022	0.015	0.397

[1] Standard error of subscore reliability estimates
[2] Mean item discriminations were 0.8, 0.9, and 1.0 for Dimensions 1, 2, and 3, respectively

that when correlations were unequal among subscores, both reliability and subscore estimates could vary across subscores. Higher reliability and more stable subscore estimates were associated with a subscore that was least correlated with other subscores. Similar results were found for a subscore whose item pool on average was most discriminating.

4 Discussion

The results demonstrate that the following factors had a positive impact on subscore reliability and the accuracy of subscore estimates: a longer subtest, a relatively larger item pool, a more discriminating item pool, a smaller variation in item difficulty, and lower subscore correlations. Among these factors, subtest length, item pool size, and item discrimination had a greater effect on subscore reliabilities and the effects of item pool size were more pronounced for a less discriminating item pool than for a more discriminating item pool.

Subscores are an important component of an MST design. Multiple factors, including test length, subscore correlations, and item pool characteristics should be taken into account when making decisions on subscores during the assessment design process. For instance, in some testing programs, when designing an MST, target subscore reliabilities may be required (e.g., a minimum of 0.70–0.75). In exploring the potential number of subscores and the minimum subtest length to support a test design, a thorough examination of the item pool is recommended.

Future research may be extended to include subtests with a mix of simple and complex structures, and item pools comprised of both dichotomous and polytomous items.

Ideally, a subscore should have a relatively high reliability estimate and be sufficiently distinct from other subscores. In addition, a subscore should have added value, that is, when predicting the same subscore on a parallel form, the subscore-based prediction is more accurate than the total-score based prediction, as illustrated by Sinharay (2013) using scores on split tests. These statistical properties of subscores also need to be investigated in MST settings in order to understand the circumstances in which subscore reporting is warranted.

References

American Psychological Association, National Council on Measurement in Education, & American Educational Research Association. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association. https://doi.org/10.1037/e578562014-008.

Diao, Q., & van der Linden, W. J. (2011). Automated test assembly using lp_solve version 5.5 in R. *Applied Psychological Measurement, 35,* 398–409. https://doi.org/10.1177/0146621610392211.

Konis, K. (2009). lpSolveAPI, version 5.5.0.15 [Computer software].

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47*(2), 150–174. https://doi.org/10.1111/j.1745-3984.2010.00106.x.

Sinharay, S. (2013). A note on assessing the added value of subscores. *Journal of Educational Measurement, 32,* 38–42. https://doi.org/10.1111/emip.12021.

Von Davier, M. (2016). *Software for multidimensional discrete latent traits models [computer software].* Princeton, NJ: Educational Testing Service.

Investigation of the Item Selection Methods in Variable-Length CD-CAT

Ya-Hui Su

Abstract Cognitive diagnostic computerized adaptive testing (CD-CAT) provides useful cognitive diagnostic information for assessment and evaluation. At present, there are only a limited numbers of previous studies investigating how to optimally assemble cognitive diagnostic tests. The cognitive discrimination index (CDI) and attribute-level discrimination index (ADI) are commonly used to select items for cognitive diagnostic tests. The CDI measures an item's overall discrimination power, and the ADI measures an item's discrimination power for a specific attribute. Su (Quantitative psychology research. Springer, Switzerland, pp. 41–53, 2018) integrated the constraint-weighted procedure with the posterior-weighted CDI and ADI for item selection in fixed-length CD-CAT, and found examinees yielded different precision. In reality, if the same precision of test results is required for all the examinees, some examinees need to take more items and some need to take fewer items than others do. To achieve the same precision for examinees, this study investigated the performance of the constraint-weighted procedure with the posterior-weighted CDI and ADI for item selection in variable-length CD-CAT through simulations.

Keywords Cognitive diagnostic computerized adaptive testing · Item selection · Constraint-weighted procedure · Variable-length

1 Introduction

Cognitive diagnostic models (CDMs) assume the latent trait to be discrete cognitive patterns, which describes if examinees have mastered or have not mastered specific skills. Many CDMs have been proposed to obtain diagnostic information (Hartz, 2002; Junker & Sijtsma, 2001; Mislevy, Almond, Yan, & Steinberg, 2000; Rupp, Templin, & Henson, 2010; Tatsuoka, 1983). One of the major CDMs applications is to implement CDMs through computerized adaptive testing (CAT), denoted as

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

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_13

cognitive diagnostic CAT (CD-CAT; Cheng, 2009; Huebner, 2010). The CD-CAT approach provides useful cognitive diagnostic information measured by psychological or educational assessments, and has been attracting a lot of practitioners' attention (Wang, Chang, & Douglas, 2012). However, there are only a limited numbers of previous studies investigating how to optimally assemble cognitive diagnostic test (Kuo, Pai, & de la Torre, 2016).

The cognitive discrimination index (CDI; Henson & Douglas, 2005) and attribute-level discrimination index (ADI; Henson, Roussos, Douglas, & He, 2008) are commonly used for item selection in CDMs. The CDI measures an item's overall discrimination power, which is the pattern-level information. By contrast, the ADI measures an item's discrimination power for a specific attribute, which is the attribute-level information. Zheng and Chang (2016) extended these two indices for item selection in CD-CAT, denoted as the posterior-weighted cognitive discrimination index (PWCDI) and posterior-weighted attribute-level discrimination index (PWADI). In their study, the PWCDI and PWADI obtained results comparable with or better than the mutual information (MI; Wang, 2013) and posterior-weighted Kullback-Leibler (PWKL; Cheng, 2009) in both short and long tests, and their computational time was shorter than that for the PWKL. They also suggested that these indices could be integrated with constraint-weighted procedures for test construction.

The priority index approach (PI; Cheng & Chang, 2009; Cheng, Chang, Douglas, & Guo, 2009) is one of the popular constraint-weighted procedures. A series of constraints are specified to include items for test construction (Stocking & Swanson, 1993; Swanson & Stocking, 1993). These constraints can be both statistical (such as target item or test information) and non-statistical (such as content specifications or key balancing) on item properties. The PI approach can manage many constraints simultaneously well. Su (2018) integrated the PI approach with the PWCDI and PWADI, denoted as the constraint-weighted PWCDI (CW-PWCDI) and constraint-weighted PWADI (CW-PWADI), for item selection in fixed-length CD-CAT. It was found that the CW-PWCDI and CW-PWADI performed slightly better than the PWCDI and PWADI in terms of attribute correct classification rates and pattern correct classification rates. In practice, if the same precision of test results is required for all the examinees, some examinees need to take more items and some need to take fewer items than others do. The fixed-length CD-CAT obtains different precision, which results in a high misclassification rate; therefore, the variable-length CD-CAT is more desirable, which achieves the required precision. However, there are problems with using precision rules as stopping rules. On one hand, some examinees may be administered undesirable lengthy tests because the required precision cannot be met. On the other hand, some examinees may be stopped too early when administering few items might have improved the precision significantly. Achieving the required precision is the main goal for a test; this depends not only on the quality of the item pool but also on the item selection procedures. To achieve the same precision for examinees, this study investigated the performance of the CW-PWCDI and CW-PWADI in variable-length CD-CAT through simulations.

1.1 Using the CW-PWCDI and CW-PWADI for Item Selection

The CDI was proposed by Henson and Douglas (2005) for test construction in CDMs. To extend Chang and Ying's (1996) Kullback–Leibler information, the CDI of item j for any two different cognitive patterns α_u and α_v is defined as follows:

$$\mathrm{CDI}_j = \frac{\sum\limits_{u \neq v} \left[h(\alpha_u, \alpha_v)^{-1} D_{juv} \right]}{\sum\limits_{u \neq v} h(\alpha_u, \alpha_v)^{-1}}, \tag{1}$$

where

$$h(\alpha_u, \alpha_v) = \sum_{k=1}^{K} (\alpha_{uk} - \alpha_{vk})^2, \tag{2}$$

and

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

α_u and α_v are $1 \times K$ attribute vectors. $P_{\alpha_u}(1)$ and $P_{\alpha_u}(0)$ are the probabilities of a correct response and an incorrect response for a given α_u, respectively, and $P_{\alpha_v}(1)$ and $P_{\alpha_v}(0)$ are the corresponding probabilities for a given α_v. X_j is the response of item j. An item with the largest CDI_j will be selected.

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

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

For item j, d_{jk1} represents the power to discriminate masters from non-masters on attribute k whereas d_{jk0} represents the power to discriminate non-masters from masters on attribute k. An item with the largest ADI_j will be selected.

To add posterior information to the CDI and ADI indices, Zheng and Chang (2016) proposed the posterior-weighted version of 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:

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

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.

To manage many constraints simultaneously, the PI approach was proposed (Cheng & Chang, 2009; Cheng et al., 2009) for item selection. The index for item j can be defined as:

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

where K is the total number of constraints, and I_j is the Fisher information of item j evaluated at the current $\hat{\theta}$. In the current study, the Fisher information is replaced with the PWCDI in Eq. (6) or the PWADI in Eq. (7) when the CW-PWCDI or CW-PWADI method is applied for item selection, respectively. The c_{jk} is 1 when the constraint k is relevant to item j; otherwise, the c_{jk} is 0. Each constraint k is associated with a weight w_k. The f_k measures the scaled ‘quota left’ of constraint k. For a content constraint k, after x_k items have been selected from the content area, the PI can be defined as:

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

For every available item in the pool $\left(c_{jk} = 0\right)$, the PI can be computed according to Eq. (8). An item with the largest value in Eq. (8) will be chosen to administer.

For item exposure control, assume constraint k requires that the item exposure rates of all items to be lower than or equal to $r\text{max}$, the f_k can be defined as:

$$f_k = \frac{1}{r\text{max}} \left(r\text{max} - \frac{n}{N} \right), \tag{10}$$

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

2 Method

In this study, the deterministic input, noisy and gate (DINA) model (Haertel, 1989; Junker & Sijtsma, 2001) was used for data generation. 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

$$P(X_{ij} = 1|s_j, g_j, \eta_{ij}) = (1 - s_j)^{\eta_{ij}} g_j^{(1-\eta_{ij})}, \tag{11}$$

where

$$\eta_{ij} = \prod\nolimits_{k=1}^{K} \alpha_{ik}^{q_{jk}}. \tag{12}$$

The η_{ij} represents if examinee i has mastered all the required attributes of item j. The s_j is the slip parameter, which measures the probability that an examinee with all the required attributes misses to obtain a correct answer for item j. The g_j is the guessing parameter, which measures the probability that an examinee without all the required attributes obtain a correct answer for item j.

A variable-length CD-CAT simulation study was carried out to evaluate the efficiency of the item selection methods. For the comparison, data generation was similar to the studies of Cheng (2009) and Zheng and Chang (2016). A total number of 500 five-attribute items were generated for the study. A total number of 3000 examinees were generated assuming every examinee has a 50% chance of mastering each attribute. The Q-matrix used in this study was generated that each item has a 30% chance of measuring each attribute. The item parameters s_j and g_j were generated from $U(0.05, 0.25)$, which represented high quality items. Three constraints considered in this study were item exposure ($\mathrm{r}_{\max} = 0.2$), content balance, and key balancing. Two factors were manipulated in this study: item selection method (CW-PWCDI, CW-PWADI, PWCDI, and PWADI) and the stopping rule (0.7, 0.8, and 0.9). When the probability of the cognitive pattern with the largest probability reaches a pre-specified value, which are 0.7, 0.8, or 0.9 in the study, a variable-length test is stop (Tatsuoka & Ferguson, 2003). To evaluate the efficiency of each method, the test length was used as a measure in a variable-length test. Besides, the constraint management and item exposure for four methods were also reported in this study.

3 Results

To evaluate the efficiency of each method, the test length, constraint management, and item exposure for four methods was reported. With respect to test length, the descriptive statistics, including maximum, minimum, mean, and standard deviation (SD), of the test length for the generated CD-CAT tests were summarized in Table 1. It was found that the constraint-weighted version (CW-PWCDI and CW-PWADI)

Table 1 The descriptive statistics of the test length for the variable-length tests

Item selection methods	Stopping rule	Maximum	Minimum	Mean	SD
CW-PWCDI	0.7	12	10	11.21	0.86
	0.8	19	15	17.33	1.12
	0.9	21	17	19.53	1.23
CW-PWADI	0.7	14	12	13.38	0.79
	0.8	21	18	19.75	1.21
	0.9	23	19	21.55	1.19
PWCDI	0.7	20	18	19.57	1.19
	0.8	23	20	21.78	1.25
	0.9	25	22	23.66	1.33
PWADI	0.7	26	23	25.57	1.25
	0.8	30	25	27.78	1.31
	0.9	37	32	34.66	1.66

tended to use fewer items than the non-constraint-weighted version (PWCDI and PWADI) while maintaining similar precision. The CW-PWCDI and CW-PWADI needed items between 11 and 22 whereas the PWCDI and CPWADI needed items between 20 and 35. This is because constraint-weighted methods could have fewer items and balance these items selected from different attributes to achieve the required precision. It was also found that the CDI-based procedures performed slightly better than the ADI-based procedures because the CDI-based procedures tended to need fewer items to achieve the same precision. That is, the CW-PWCDI performed better than the CW-PWADI, and the PWCDI performed better than the PWADI. When the required precision was high (e.g., 0.9), more items were needed for all procedures.

With respect to the constraint management and item exposure, the number of averaged violations and the number of item overexposed for all procedures were reported in Table 2. It was found that the constraint-weighted version (CW-PWCDI and CW-PWADI) outperformed the non-constraint-weighted version (PWCDI and PWADI) in terms of no violations and no item overexposed. The PWADI performed better than the PWCDI since the PWADI has less violation and less item overexposed. The PWADI obtained the averaged violations between 1.2 and 2.2, and overexposed items between 3 and 8. The PWCDI obtained the averaged violations between 1.6 and 2.6, and overexposed items between 6 and 12. When the required precision was high (e.g., 0.9), higher averaged violations and more overexposed items were found for the PWCDI and PWADI.

Table 2 The number of averaged violations and the number of item overexposed for the variable-length test

Item selection methods	Stopping rule	#Violations	#Item overexposed
CW-PWCDI	0.7	0	0
	0.8	0	0
	0.9	0	0
CW-PWADI	0.7	0	0
	0.8	0	0
	0.9	0	0
PWCDI	0.7	1.6	6
	0.8	2.2	8
	0.9	2.6	12
PWADI	0.7	1.2	3
	0.8	1.9	5
	0.9	2.2	8

4 Discussions

The CD-CAT provides useful cognitive diagnostic information for teachers, parents, and students in psychological or educational assessments. This study integrated the constraint-weighted procedure with the PWCDI and PWADI for item selection in variable-length CD-CAT. It was found that the CW-PWCDI and CW-PWADI outperformed the PWCDI and PWADI in terms of test length (short test length), constraint management (no constraint violation), and item exposure (no item overexposed). Besides, the CW-PWCDI and CW-PWADI can be implemented easily and computed efficiently. The constraint-weighted item selection procedures (i.e. the CW-PWCDI and CW-PWADI) have great potential for item selection in operational CD-CAT.

Some future research lines are addressed as follows. In practice, the performance of item selection methods depends on the quality and structure of the item pool. This study only considered the simulated item bank with five-attribute DINA model, which was similar to previous studies (Cheng, 2009; Zheng & Chang, 2016). Besides, constraints on item exposure, content balance, and key balancing were considered in the study. It would be worth to investigate the efficiency of item selection methods in an operational CD-CAT pool with different attributes, different constraints, other CDMs, and other constraint-weighted procedures.

References

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.

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) Champaign, IL: University of Illinois at Urbana-Champaign.

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 & 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.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277–292.

Su, Y.-H. (2018). Investigating the constrained-weighted item selection methods for CD-CAT. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative Psychology Research* (Vol. 233, pp. 41–53). Switzerland: Springer. https://doi.org/10.1007/978-3-319-77249-3_4.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17,* 151–166.

Tatsuoka, C., & Ferguson, T. (2003). Sequential classification on partially ordered sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65,* 143–157.

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

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement, 73,* 1017–1035.

Wang, C., Chang, H. H., & Douglas, J. (2012). Combining CAT with cognitive diagnosis: A weighted item selection approach. *Behavior Research Methods, 44,* 95–109.

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

A Copula Model for Residual Dependency in DINA Model

Zhihui Fu, Ya-Hui Su and Jian Tao

Abstract Cognitive diagnosis models (CDMs) have been received the increasing attention by educational and psychological assessment. In practice, most CDMs are not robust to violations of local item independence. Many approaches have been proposed to deal with the local item dependence (LID), such as conditioning on other responses and additional random effects (Hansen In Hierarchical item response models for cognitive diagnosis. University of California, LA, 2013); however, these have some drawbacks, such as non-reproducibility of marginal probabilities and interpretation problem. (Braeken et al. In Psychometrika 72(3): 393–411 2007) introduced a new class of marginal models that makes use of copula functions to capture the residual dependence in item response models. In this paper, we applied the copula methodology to model the item dependencies in DINA model. It is shown that the proposed copula model could overcome some of the dependency problems in CDMs, and the estimated model parameters recovered well through simulations. Furthermore, we have extended the R package *CDM* to fit the proposed copula DINA model.

Keywords Cognitive diagnosis models · Copula model · Local item dependence

Z. Fu (✉)
Department of Statistics, School of Mathematics and System Science,
Shenyang Normal University, Shenyang 110034,
Liaoning, People's Republic of China
e-mail: fuzhihui2001@163.com

Y.-H. Su
Department of Psychology, National Chung Cheng University,
168 University Road, Minhsiung Township, Chiayi County 62102, Taiwan

J. Tao
Key Laboratory of Applied Statistics of MOE,
School of Mathematics and Statistics, Northeast Normal University,
Changchun 130024, Jilin, People's Republic of China

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_14

1 Introduction

The deterministic inputs, "noisy" and"gate" (DINA) model (Junker, & Sijtsma, 2001) is a popular conjunctive cognitive diagnosis models (CDMs) which assumes that a respondent must have mastered all required attributes in order to correctly respond to an item on an assessment. To estimate respondent's knowledge of attributes, we need information about which attributes are required for each item. For this, we use a Q-matrix which is an $J \times K$ matrix where $q_{jk} = 1$ if item j requires attribute k and 0 if not. I is the number of items and K is the number of attributes in the assessment. A binary latent variable α_{ik} indicates respondent i's knowledge of attribute k, where $\alpha_{ik} = 1$ if respondent i has mastered attribute k and 0 if he or she has not. Then, an underlying attribute profile of respondent i, $\boldsymbol{\alpha}_i$, is a binary vector of length K that indicates whether or not the respondent has mastered each of the K attributes.
The deterministic element of the DINA model is a latent variable η_{ij} that indicates whether or not respondent i has mastered all attributes required for item j:

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

for $i = 1, 2, \cdots, I$, I denotes the number of examinees; $j = 1, 2, \cdots, J$ and $k = 1, 2, \cdots, K$. If respondent i has mastered all attributes required for item j, $\eta_{ij} = 1$; if the respondent has not mastered all of the attributes, $\eta_{ij} = 0$.

$$s_j = P(Y_{ij} = 0|\eta_{ij} = 1), g_j = P(Y_{ij} = 0|\eta_{ij} = 0)$$

The slip parameter s_j is the probability that respondent i responds incorrectly to item j although he or she has mastered all required attributes. The guess parameter g_j is the probability that respondent i responds correctly to item j although he or she has not mastered all the required attributes. It follows that the probability of a correct response of respondent i to item j with skill vector $\boldsymbol{\alpha}_i$ is

$$P_j(\boldsymbol{\alpha}_i) = P(Y_{ij} = 1|\boldsymbol{\alpha}_i) = g_j^{1-\eta_{ij}}(1 - s_j)^{\eta_{ij}} \tag{2}$$

A basic assumption of the DINA model is local conditional independence:

$$P(\boldsymbol{Y}_i = \boldsymbol{y}_i) = \prod_{j=1}^{J} P(Y_{ij} = y_{ij}|\boldsymbol{\alpha}_i) = \prod_{j=1}^{J} P_j(\boldsymbol{\alpha}_i)^{y_{ij}}[1 - P_j(\boldsymbol{\alpha}_i)]^{(1-y_{ij})} \tag{3}$$

where $\boldsymbol{Y}_i$ is the response vector of examinee i on the set of J items. The DINA model and most CDMs are not robust to violations of local item independence. Potential causes of dependence could be from the fact that the model is too simple, therefore more general CDMs (e.g., the LCDM and GDINA) should be considered for some practices (de la Torre, 2011; Henson, Templin, & Willse, 2009; von, 2008). Residual item dependence might also be interpreted as the Q-matrix misspecification (That is, a failure to include all relevant attributes in the model, see (Chen, Culpepper, Chen, & Douglas, 2018; Culpepper and Chen, 2018; Chen, Liu, Xu, & Ying, 2015; Xu, & Shang, 2018). Local item independence (LID) can affect the estimation and the reliability of the model parameters. Ignoring such dependencies, i.e., using a traditional CDM that assumes local independency may affect the estimates of model parameters and misclassification of respondents.

Hansen (2013) have included additional random effects to account for the dependency; this approach is followed by testlet models (Bradow, 1999), random intercept item factor model (Maydeu-Olivares, & Coffman, 2006). To deal with LID, (Braeken, Tuerlinckx, & De Boeck, 2007) primitively introduced a convenient marginal copula modeling tool to construct multivariate distributions for dependent responses in IRT models. In this paper, we will also employ copula-based technique to model the dependencies in CDMs. The remaining manuscript is arranged as follows.

In Sect. 2, we describe the some basic theory of copula function and present the proposed Copula CDMs as an extension. In Sect. 3, we present the EM algorithm to estimate the marginal and associates parameters. In Sect. 4, Simulation studies are demonstrated to show the consequences of local item independence violations within the context of cognitive diagnostic modeling. Finally, we conclude with a few summary remarks.

2 Modeling

2.1 An Overview of Copula Theory

In mathematics, a copula function can be defined as a R-dimensional distribution function $C : [0, 1]^R \to [0, 1]$ that relates a multivariate uniform cumulative distribution function (CDF) to its univariate margins that have uniform CDFs (Nelsen, 2006). The Sklar's theorem (Sklar, 1959) ensures that for any R-dimensional distribution function F_Y with univariate marginal $F_{Y_1}, F_{Y_2}, \cdots, F_{Y_R}$, there exist a copula function C such that

$$F_Y(y_1, \cdots, y_R) = C(F_{Y_1}(y_1), \cdots, F_{Y_R}(y_R))$$

Two most widely used copula families are elliptical copulas and Archimedean copulas. In this paper, we focused on the Archimedean copulas which have a simple structure and can be written as:

$$C(u_1, \cdots, u_R) = \phi^{-1}[\phi(u_1) + \cdots + \phi(u_R)]$$

2.2 Copula Model for DINA Model

In this section, we extend Braeken et al. (2007)'s copula-based joint approach to the setting of CDMs. Specifically, we consider S disjoint subsets of $\{1, \cdots, J\}$ denoted as $J_1, \cdots, J_s, \cdots, J_S$, where J_s has cardinality I_s. The response vector of examinee i, $\boldsymbol{Y}_i$ is similarly divided into subsets $\boldsymbol{Y}_i^{(1)}, \cdots, \boldsymbol{Y}_i^{(s)}, \cdots, \boldsymbol{Y}_i^{(S)}$, where $\boldsymbol{Y}_i^{(s)} = (Y_{ij}^{(s)}, j \in J_s)$, $s = 1, 2, \cdots, S$. The different subsets are assumed independent.

$$\begin{aligned} P(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{\alpha}_i) &= \prod_{s=1}^{S} P(\boldsymbol{Y}_i^{(s)} = \boldsymbol{y}_i^{(s)}|\boldsymbol{\alpha}_i) \\ &= \prod_{s=1}^{S} P(Y_{ij}^{(s)} = y_{ij}^{(s)}, j \in J_s|\boldsymbol{\alpha}_i), i = 1, 2, \cdots, I. \end{aligned} \tag{4}$$

(**Note** the superscript (s) in $Y_{ij}^{(s)}$ and $y_{ij}^{(s)}$ are dropped in the remainder of this article for simplicity.) Let $b_{ij} = F_{ij}(y_{ij})$, and $a_{ij} = F_{ij}(y_{ij}^{-})$ be the left hand limit of marginal distribution function F_{ij} at y_{ij}, with $F_{ij}(0^-) = 0$, $F_{ij}(1) = 1$, and $F_{ij}(0) = F_{ij}(1^-) = 1 - P_j(\boldsymbol{\alpha}_i) = s_j^{\eta_{ij}}(1 - g_j)^{1-\eta_{ij}}$ (see Eq. (2)).
For responses in the subset s, the joint probability is evaluated from a copula function C_s

$$\begin{aligned} &P(Y_{ij} = y_{ij}, j \in J_s|\boldsymbol{\alpha}_i) \\ &= P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \cdots, Y_{iI_s} = y_{iI_s}|\boldsymbol{\alpha}_i) \\ &= \Delta_{a_{i1}}^{b_{i1}} \Delta_{a_{i2}}^{b_{i2}} \cdots \Delta_{a_{iI_s}}^{b_{iI_s}} C_s(\boldsymbol{v}_i) \end{aligned} \tag{5}$$

where $\boldsymbol{v}_i = (v_{i1}, \cdots, v_{iI_s})$ and we employ the difference notation of (Nelsen, 2006):

$$\Delta_{a_{i1}}^{b_{i1}} C_s(u_{i1}, \cdots, u_{i,j-1}, v_{ij}, u_{i,j+1}, \cdots, u_{iI_s}) =$$

$$C_s(u_{i1}, \cdots, u_{i,j-1}, b_{ij}, u_{i,j+1}, \cdots, u_{iI_s}) - C_s(u_{i1}, \cdots, u_{i,j-1}, a_{ij}, u_{i,j+1}, \cdots, u_{iI_s})$$

where v_{ij} is an index of differencing. For example, when $I_s = 3$,

$$\begin{aligned} P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, Y_{i3} = y_{i3}) &= \Delta_{a_{i1}}^{b_{i1}} \Delta_{a_{i2}}^{b_{i2}} \Delta_{a_{i3}}^{b_{i3}} C_s(v_{i1}, v_{i2}, v_{i3}) \\ = C_s(b_{i1}, b_{i2}, b_{i3}) - C_s(b_{i1}, b_{i2}, a_{i3}) &- C_s(b_{i1}, a_{i2}, b_{i3}) - C_s(a_{i1}, b_{i2}, b_{i3}) \\ +C_s(b_{i1}, a_{i2}, a_{i3}) + C_s(a_{i1}, b_{i2}, a_{i3}) &+ C_s(a_{i1}, a_{i2}, b_{i3}) - C_s(a_{i1}, a_{i2}, a_{i3}) \end{aligned}$$

The regular conditional independent DINA model arises as a special case when $S = 1$ and C_s is the independence copula

$$\prod(F_{Y_{i1}|\boldsymbol{\alpha}_i}, \cdots, F_{Y_{iJ}|\boldsymbol{\alpha}_i}) = \prod\nolimits_{j=1}^{J} F_{Y_{ij}|\boldsymbol{\alpha}_i}.$$

The regular conditional independent DINA model arises as a special case when $S = 1$ and C_s is the independence copula:

$$\prod(F_{Y_{i1}|\boldsymbol{\alpha}_i}, \cdots, F_{Y_{iJ}|\boldsymbol{\alpha}_i}) = \prod\nolimits_{j=1}^{J} F_{Y_{ij}|\boldsymbol{\alpha}_i}$$

3 Parameter Estimation for Copula CDMs

Parameter estimation of the Copula DINA model is performed by marginal maximum likelihood (MML) estimation method and implemented using the expectation-maximization (EM; see (Dempster, Laird, & Rubin, 1977)) algorithm. The marginal likelihood for the Copula DINA model is

$$L(\boldsymbol{Y}) = \prod_{i=1}^{I} L(\boldsymbol{Y}_i) = \prod_{i=1}^{I}\sum_{l=1}^{L} L(\boldsymbol{Y}_i|\boldsymbol{\alpha}_l)P(\boldsymbol{\alpha}_l) \tag{6}$$

The conditional likelihood of $\boldsymbol{Y}_i$ can be written as

$$L(\boldsymbol{Y}_i|\boldsymbol{\alpha}_l) = \prod_{s=1}^{S} P(\boldsymbol{Y}_i^{(s)}|\boldsymbol{\alpha}_l) = \prod_{s=1}^{S} P(Y_{ij} = y_{ij}, j \in J_s|\boldsymbol{\alpha}_l) \tag{7}$$

E-step

a. The individual posterior distribution for the skills $\boldsymbol{\alpha}_l$, $l = 1, \ldots, L$

$$P(\boldsymbol{\alpha}_l|\boldsymbol{Y}_i) = \frac{P(\boldsymbol{Y}_i|\boldsymbol{\alpha}_l)P(\boldsymbol{\alpha}_l)}{\sum_{m=1}^{L} P(\boldsymbol{Y}_i|\boldsymbol{\alpha}_m)P(\boldsymbol{\alpha}_m)} \tag{8}$$

b. Two types of expected numbers are obtained from the posterior: The first count

$$T_{jl} = \sum_{i=1}^{I} P(\boldsymbol{\alpha}_l|\boldsymbol{Y}_i)$$

$$R_{jl} = \sum_{i=1}^{I} \boldsymbol{Y}_i P(\boldsymbol{\alpha}_l|\boldsymbol{Y}_i)$$

denotes the expected number of examinees which are classified into skill class $\boldsymbol{\alpha}_l$ for item j, $j = 1, \ldots, J$. The second count denotes the expected number of examinees classified in skill class $\boldsymbol{\alpha}_l$ while answering item j correctly.

M-step
a. Update $\hat{g}_j, \hat{s}_j$

Independent case

$$\frac{\partial \log L(\mathbf{Y})}{\partial \boldsymbol{\beta}_j} = \sum_{i=1}^{I} P(\boldsymbol{\alpha}_l|\boldsymbol{Y}_i) \sum_{l=1}^{L} \frac{\partial \log P_j(\boldsymbol{\alpha}_i)^{y_{ij}}[1 - P_j(\boldsymbol{\alpha}_i)]^{(1-y_{ij})}}{\partial \boldsymbol{\beta}_j} = 0$$

Then the derived item estimators have the same closed form of those issued from the traditional DINA model:

$$\hat{g}_j = \frac{R_j^{(0)}}{T_j^{(0)}} \text{ and } \hat{s}_j = 1 - \frac{R_j^{(1)}}{T_j^{(1)}} \tag{9}$$

where

$$T_j^{(0)} = \sum\nolimits_{l=1}^{L} T_{jl}(1 - \eta_{lj}), R_j^{(0)} = \sum\nolimits_{l=1}^{L} Y_{ij}T_{jl}(1 - \eta_{lj}),$$
$$T_j^{(1)} = \sum\nolimits_{l=1}^{L} T_{jl}\eta_{lj}, R_j^{(1)} = \sum\nolimits_{l=1}^{L} Y_{ij}T_{jl}\eta_{lj}$$

dependent case

For items exhibit some dependence and form a dependent subset, we have

$$\frac{\partial \log L(\mathbf{Y})}{\partial \boldsymbol{\beta}_j} = \sum_{i=1}^{I} P(\boldsymbol{\alpha}_l|\boldsymbol{Y}_i) \sum_{l=1}^{L} \frac{\partial \log P(Y_{ij} = y_{ij}, j \in J_s|\boldsymbol{\alpha}_l)}{\partial \boldsymbol{\beta}_j} = 0$$

No closed-form solutions of this equation exist. An iterative algorithm was used to find the optimal solution.
b. Update $P(\boldsymbol{\alpha}_l)$

$$P(\boldsymbol{\alpha}_l) = \sum\nolimits_{i=1}^{I} P(\boldsymbol{\alpha}_l|Y_i)/I \tag{10}$$

and the skill mastery probabilities are defined as

$$P(\alpha_k) = \sum\nolimits_{l=1}^{L} \alpha_{lk} P(\boldsymbol{\alpha}_l|\boldsymbol{Y}_i) \tag{11}$$

The E- and M-Step alternate until convergence. EM algorithm was implemented by R and we extend the *CDM* package to estimate Copula DINA/DINO.

4 Illustrated Examples

4.1 Design

We compare the results of correctly specified model-Copula CDM (which used to generate the data), with those issued by the misspecified model-traditional CDM. Detailed design are listed in Table 1.

4.2 Results

To save space, we only list partial results. And detailed results can be requested from the author.

Part 1 Comparison of item parameter estimators Fig. 1 illustrated the estimated guessing and slipping parameter issued from the DINA model and the Copula-DINA model for $J = 20, I = 500/1000/2000$. By observing the distance from the points to the line, we find that the Copula-DINA estimates are much closer to the true values, compared with the DINA estimates which deviated from the true line.

Part 2 Bias and RMSE Table 2 gives gives Bias and RMSE of the guessing, slipping, and skill class parameters. The results indicate that the Copula-DINA model provides accurate parameter estimates compared with DINA model. Under the same conditions, all RMSE values derived from Copula-DINA model are smaller than those computed from the DINA model. Further, for each model, the RMSE decreases with increasing sample size.

Table 1 Simulation Design

Condition	Specification
number of examinees	I = 500, 1000, or 2000
number of items	J = 20 or 30
number of attributes	K = 5
dependent subsets	$J_1 = \{1, 2\} J_2 = \{5, 6\}, J_3 = \{10, 11, 12\}$
Model	Copula-DINA versus DINA or Copula-DINO versus DINO
Copula function	Frank's copula, Gumbel's copula
Copula dependency parameter	$[\delta_{J_1} = 2, \delta_{J_2} = 4, \delta_{J_3} = 6]$ or $[\delta_{J_1} = 4, \delta_{J_2} = 6, \delta_{J_3} = 8]$
True value:	$s = g = 0.2$
Replications:	$R = 100$

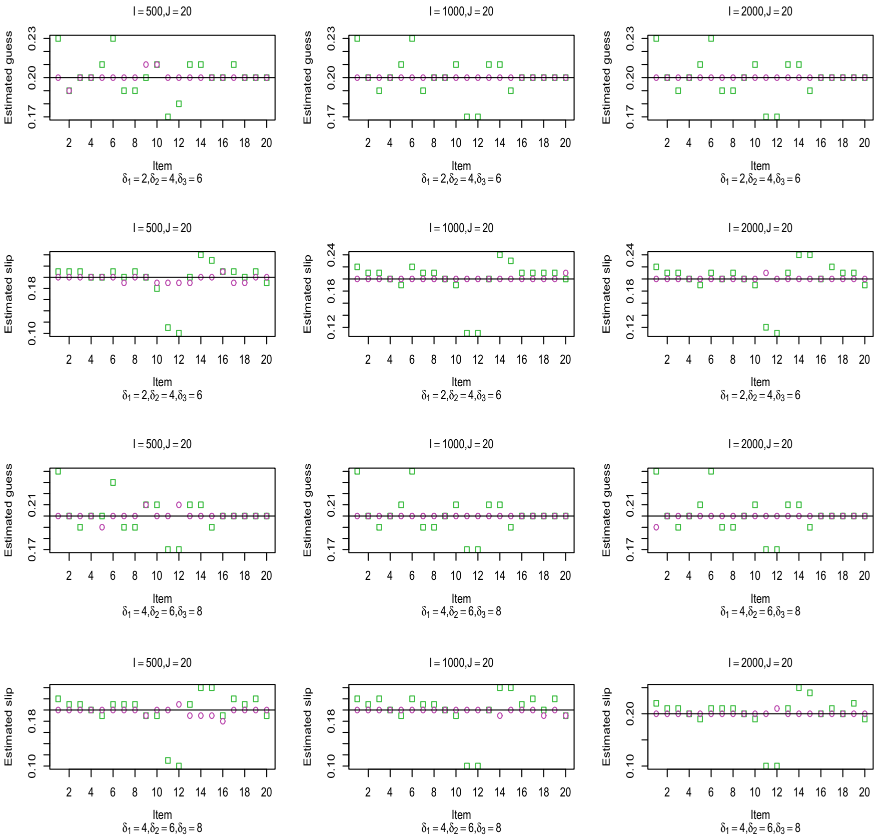


Fig. 1 Comparison of estimated guess and slip parameters issued from the DINA model and the Copula-DINA model, test length $J = 20$, total number of examinees $I = 500/1000/2000$. Note: solid line denotes the true value, □ denotes estimates from DINA, and ○ represents the estimates from Copula-DINA

Part 3 Difference of Person Classifications The IRT.factor.scores are computed by reference to the R package "CDM" to compare the individual classifications obtained from the Copula-CDM and CDM. First, we computed the classifications for each model using the function IRT.factor.scores:

CDM::IRT.factor.scores(DINA, type = "MLE")= class-independent
CDM::IRT.factor.scores(Copula-DINA, type = "MLE")= class-copula

Second, count the number of different classifications between the above two types of models, and divided by the total number of examinees I, then got the ratio of classification differences (RCD) which listed in Table 3.

$$\texttt{RCD} = \sum (\text{class-independent} \neq \text{ class-copula})/I$$

Table 2 Comparison of estimated results issued from the DINA model and the Copula-DINA model, test length is $J = 20$, total number of examinees $I = 500/1000/2000$; dependency: $\delta_{J_1} = 4$, $\delta_{J_2} = 6$, $\delta_{J_3} = 8$

			Bias			RMSE		
I	Parameter	model	mean	min	max	mean	min	max
500	g_j	DINA	0.0003	−0.0345	0.0421	0.0359	0.0214	0.0586
		Copula-DINA	0.0002	−0.0066	0.0068	0.0275	0.0191	0.0394
	s_j	DINA	−0.0017	−0.0998	0.0401	0.056	0.0319	0.1175
		Copula-DINA	−0.0025	−0.0178	0.0063	0.0393	0.0275	0.0567
	$p(\alpha_l)$	DINA	0	−0.0124	0.012	0.0139	0.0102	0.0218
		Copula-DINA	0	−0.0023	0.0022	0.0116	0.0094	0.0162
1000	g_j	DINA	0.0009	−0.0322	0.0421	0.0264	0.0149	0.0522
		Copula-DINA	0.0003	−0.0036	0.0036	0.0191	0.0138	0.0265
	s_j	DINA	−0.0019	−0.1038	0.0404	0.0418	0.0218	0.1117
		Copula-DINA	−0.0014	−0.0088	0.0042	0.0272	0.0201	0.0389
	$p(\alpha_l)$	DINA	0	−0.0127	0.0122	0.0105	0.0066	0.0176
		Copula-DINA	0	−0.002	0.0018	0.0081	0.0062	0.0107
2000	g_j	DINA	0.0011	−0.0328	0.0385	0.0208	0.0115	0.0439
		Copula-DINA	−0.0003	−0.0061	0.0045	0.0136	0.0111	0.0174
	s_j	DINA	−0.0012	−0.1043	0.0482	0.035	0.016	0.1093
		Copula-DINA	−0.0003	−0.0038	0.0058	0.0191	0.0138	0.0258
	$p(\alpha_l)$	DINA	0	−0.014	0.0126	0.0083	0.0052	0.0149
		Copula-DINA	0	−0.001	0.0014	0.0056	0.0047	0.0071

It is easy to see that the ratio of differences increase with the dependency δ. From Table 3, we can conclude that the classification differences can not be ignored for the current design.

Part 4 Assessing Model Fit and Local Dependence We computes several measures of absolute model fit and local dependence indices for dichotomous item responses which are based on comparing observed and expected frequencies of item pairs (see (Chen, de la Torre, & Zhang, 2013) for details). We extend the R package function *CDM::modelfit.cor.din* and *CDM::IRT.modelfit* to the Copula-CDM model.
For each fit statistics, it holds that smaller values (values near to zero) indicate better fit. These indexes were computed from both the traditional and copula CDMs over

Table 3 Ratio of classification difference (RCD) under 24 conditions

		DINA versus Copula-DINA		DINO versus Copula-DINO	
J	I	$\delta_{J_1}, \delta_{J_2}, \delta_{J_3}$ = 2,4,6	$\delta_{J_1}, \delta_{J_2}, \delta_{J_3}$ = 4,6,8	$\delta_{J_1}, \delta_{J_2}, \delta_{J_3}$ = 2,4,6	$\delta_{J_1}, \delta_{J_2}, \delta_{J_3}$ = 4,6,8
		RCD	RCD	RCD	RCD
20	500	89/500	99/500	90/500	95/500
	1000	182/1000	197/1000	180/1000	193/1000
	2000	380/2000	403/2000	377/2000	399/2000
30	500	57/500	60/500	53/500	59/500
	1000	113/1000	123/1000	108/1000	120/1000
	2000	236/2000	255/2000	227/2000	246/2000

the range of simulation conditions. Results from these analyses were presented in Table 4. On the basis of all the listed indexes, the results from Copula DINA/DINO model are consistently smaller than those from DINA/DINO model.

Test of global absolute model fit
In Table 5, the statistic max(X2) denotes the maximum of all $\chi^2_{jj'}$ statistics accompanied with a p value. A similar statistic abs(fcor) is created as the absolute value of the deviations of Fisher transformed correlations as used in (Chen, de la Torre, & Zhang, 2013). The index demonstrated high rate of acceptance for copula model.

$$\chi^2_{jj'} = \sum_{n=0}^{1} \sum_{m=0}^{1} \frac{\left[N(Y_j = n, Y_{j'} = m) - \hat{P}(Y_j = n, Y_{j'} = m)\right]}{\hat{P}(Y_j = n, Y_{j'} = m)}$$

5 Conclusions and Future Study

The purpose of this research was to apply the copula methodology to model the item dependencies in CDMs. There are two advantages of copula modeling method: first, the marginal model and dependency part can be modeled separately; second, it is not restricted to model a linear dependency structure as implied by the use of the multivariate normal distribution.

In the simulation study, we assess the performance of proposed copula model and estimating methods, where we obtain the satisfactory results under the current conditions. The results show that the parameters are recovered and can be estimated properly with minor bias. Otherwise, results from simulation demonstrate that failure to account for dependency can result biased item parameter estimates, misclassification of examinees.

For future studies: Firstly, more simulation conditions should be considered. Secondly, some complex diagnostic model(e.g., Gdina model) should be considered.

Table 4 Model fit statistics results under dependency $\delta_{J_1} = 4, \delta_{J_2} = 6, \delta_{J_3} = 8$

J	*I*	Model	AIC	BIC	MADcor	SRMSR	MX2	100*MAD RESIDCOV	MADQ3	MADaQ3
20	500	DINA	12347	12647	0.039	0.065	2.105	0.937	0.065	0.062
		Copula-DINA	11963	12262	0.031	0.038	0.781	0.727	0.061	0.059
	1000	DINA	24624	24972	0.031	0.059	3.476	0.744	0.059	0.056
		Copula-DINA	24053	24401	0.022	0.027	0.775	0.514	0.054	0.052
	2000	DINA	49180	49578	0.026	0.056	6.173	0.616	0.056	0.053
		Copula-DINA	48740	49138	0.015	0.019	0.765	0.36	0.049	0.047
	500	DINO	12341	12640	0.038	0.063	1.966	0.914	0.064	0.061
		Copula-DINO	12005	12305	0.03	0.038	0.761	0.716	0.06	0.058
	1000	DINO	24634	24983	0.031	0.057	3.265	0.736	0.058	0.055
		Copula-DINO	24152	24500	0.022	0.027	0.779	0.516	0.053	0.052
	2000	DINO	49187	49585	0.025	0.054	5.765	0.607	0.054	0.052
		Copula-DINO	48878	49276	0.015	0.019	0.767	0.362	0.049	0.047

Table 5 Test of global model fit results issued from the CDM model and the Copula-CDM model

			$\delta_{s=1} = 2, \delta_{s=2} = 4, \delta_{s=3} = 6$				$\delta_{s=1} = 4, \delta_{s=2} = 6, \delta_{s=3} = 8$			
			max(X2)		abs(fcor)		max(X2)		abs(fcor)	
J	*I*	Model	value	p	value	p	value	p	value	p
20	500	DINA	85.42	0	0.5	0	90.91	0	0.52	0
		Copula-DINA	6.64	0.9	0.12	0.77	6.92	0.86	0.12	0.71
	1000	DINA	167.17	0	0.49	0	177.33	0	0.52	0
		Copula-DINA	6.68	0.89	0.08	0.75	6.64	0.9	0.08	0.75
	2000	DINA	338.46	0	0.5	0	350.34	0	0.52	0
		Copula-DINA	6.72	0.89	0.06	0.77	6.58	0.88	0.06	0.76
	500	DINO	73.88	0	0.45	0	79.76	0	0.48	0
		Copula-DINO	6.79	0.88	0.12	0.74	6.55	0.89	0.11	0.78
	1000	DINO	147.72	0	0.46	0	161.59	0	0.49	0
		Copula-DINO	6.58	0.91	0.08	0.77	6.73	0.89	0.08	0.77
	2000	DINO	289.34	0	0.45	0	313.27	0	0.48	0
		Copula-DINO	6.81	0.86	0.06	0.75	6.3	0.94	0.06	0.86

Thirdly, for dependent item sets with large cardinality I_s, optimization problem gets harder. We can combine the EM algorithm with the two-stage estimation method-IFM method (Joe, & Xu, 1996). Further corresponding simulation studies are still needed. Fourthly, model selection-we have to choose the most adequate diagnostic model for the marginal probabilities and find out which copula function shows the best representation of the dependence structure.

References

Bradow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.

Braeken, J., Tuerlinckx, F., & De Boeck, P. (2007). Copula Functions for Residual Dependency. *Psychometrika*, *72*(3), 393–411.

Chen, Y., Culpepper, S. A., Chen, Y., & Douglas, J. (2018). Bayesian Estimation of the DINA Q matrix. *Psychometrika*, *83*, 89–108.

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, *110*, 850–866.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*, 123–140.

Culpepper, S. A., & Chen, Y. (2018). Development and application of an axploratory reduced reparameterized unified model. *Journal of Educational and Behavioral Statistics*,. https://doi.org/10.3102/1076998618791306.

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

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood wiih incomplete data via the EM algorithm. Journal of the Royal Statistical Society. *39*, 1–38 (Series B).

Hansen, M. (2013). Hierarchical item response models for cognitive diagnosis. Unpublished doctoral dissertation, University of California, LA.

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.

Joe, H., & Xu, J. (1996). *The estimation method of inference functions for margins for multivariate models, technical Report 166*. Department of Statistics: University of British Columbia.

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.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*, 344–362.

Nelsen, R. B. (2006). *An introduction to copulas*. New York: Springer.

Sklar, A. W. (1959). Fonctions de répartition àn dimension et leurs marges. *Publications de lInstitut de Statistique de lUniversitéde Paris*, *8*, 229–231.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307.

Xu, G., & Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association*, *113*, 1284–1295.

A Cross-Disciplinary Look at Non-cognitive Assessments

Vanessa R. Simmreing, Lu Ou and Maria Bolsinova

Abstract The past two decades have seen an increasing interest in studying non-cognitive skills across disciplines. Despite the shared popularity, non-cognitive skills have been assessed variously across disciplines with different assumptions and target populations. Synthesizing across the commonalities, differences, and limitations in these various approaches will have important implications for the development and interpretation of non-cognitive assessments. In this project, we review the ways in which non-cognitive skills have been conceptualized and measured across psychology and education, and use self-control as an example to address the challenges to various types of assessments that are commonly seen in these disciplines. We will draw implications from a cross-disciplinary perspective on the validity and reliability of the non-cognitive assessments.

Keywords Non-cognitive · Assessment · Validity · Reliability

1 Introduction

1.1 *What Does "Non-cognitive" Mean?*

Non-cognitive skills have been an increasingly popular target of assessment, but what do we mean when we say "non-cognitive"? Easton (2013) noted "Everybody hates this term but everyone knows roughly what you mean when you use it" (p. 8; see also Duckworth & Yeager, 2015). Table 1 shows a non-exhaustive list of twelve adjectives and six nouns that are variously combined across research areas; additional specific sub-skills, such as "grit" and "growth mindset", have been popular in recent research (Kyllonen, Bertline, & Zu, 2018). A search of the Web of Science Core Collection on the intersections of these terms indicated more than 17,500 publications, with

V. R. Simmreing (✉) · L. Ou · M. Bolsinova
ACT, Inc., Iowa City, IA 52243, USA
e-mail: Vanessa.simmering@act.org

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_15

Table 1 Range of adjectives and nouns used together to refer to constructs we discuss here (from Kyllonen et al., 2018)

Adjectives		Nouns
Non-cognitive	Non-academic	Attributes
Socio-emotional	Psychosocial	Competencies
Soft	21st-Century	Skills
Personal	Intra-/inter-personal	Traits
Self-management	Character	Strengths
Meta-cognitive	Emotional	Intelligence

dramatic growth in the rate of publication in the past two decades, from less than 300 in 1998 to over 1100 in 2017.

If we "roughly" know what is meant by non-cognitive and related terms, as Easton (2013) argued, is it important to consider differences in terminology? We argue it is because the choice of terminology is more than semantics: each of these terms was developed with particular underlying assumptions and goals. For example, the choice of referring to non-cognitive "traits" versus "skills" or "competencies" implies different expectations for malleability—we typically think of traits as stable characteristics of a person, whereas skills or competencies are more likely to be developed and improve. The underlying assumptions and goals across disciplines then influence what type of measures will be proposed and investigated through research. In the current paper, we consider the history of such terms across disciplines, where the initial motivation to develop assessments and details of the goals and populations differ.

1.2 When "Non-cognitive" Appeared in the Literature

Our first question was when and how the term "non-cognitive" emerged in the social sciences literature. We limited our search to this term for simplicity, with the general principles of our argument being applicable to other terms as well (although the specific details differ). Table 2 shows some of the earliest appearance of this term across disciplines based on a Web of Science Core Collection search.

As these examples show, there were three general disciplines in which the term was presented in the early- to mid-20th century. First was social psychology, a discipline in which there was a straightforward contrast with cognitive psychology encompassing affect/emotion, personality, and interpersonal interactions. Next was education, in which the contrast was with academic or intellectual abilities, which have traditionally been the targets of assessments. Lastly this term emerged in the context of training and the workforce; in these contexts non-cognitive skills were differentiated from more typical metrics from education, and the goal was more specific differentiation and prediction of success. There is some similarity in the emergence of the term non-cognitive across disciplines, but there are also important differences.

Table 2 Sample of early publications using the term “non-cognitive”

Publication title	Year	Discipline
An experimental study of the cognitive and affective processes in retention and recall	1914	Social psychology
Personality tests—whither?	1933	Social psychology
The measurement of socially valuable qualities	1947	Social psychology
Non-intellective factors in certain intelligence and achievement tests	1951	Education
University selection: Some psychological and educational factors	1953	Education
The validity of several non-cognitive tests as predictors of certain naval office candidate school criteria	1954	Training/workforce
Differential testing of high-level personnel	1957	Training/workforce

For the purposes of this paper, we will consider the intersection of social psychology and workforce under the umbrella of industrial/organizational (I/O) psychology, and divide education into later (higher/secondary) and earlier (primary/preschool) because these align with major areas of the literature. This discussion, as well as the samples above, are not intended to be an exhaustive review of the literature, so other disciplines in which the term has been used (e.g., neuropsychology) will not be considered.

2 Comparing Across Disciplines and Assessment Types

When developing and comparing non-cognitive assessments, a number of specific challenges arise related to underlying assumptions about the constructs of interest, which may differ across disciplines of origin and assessment types. We first consider a general comparison between disciplines, then specific examples of assessments of one non-cognitive factor, self-control. Although the assessments of self-control arose within the same discipline (personality psychology), the examples we use illustrate characteristics of assessments that reflect different assumptions about the underlying constructs.

2.1 Comparing Disciplines: Who Is Assessed and for What Purpose

Historically, different disciplines developed assessments of non-cognitive skills for different purposes or goals, as Table 2 suggests, driven by different assumptions about the underlying constructs, and targeting different populations. Comparing I/O psych with education, we can see similarities in later education, when the targeted populations and goals are similar, but difference in earlier education due to the different populations who require different methods and potentially different goals. What is less clear in this comparison, however, is whether the assumptions underlying the constructs are similar in cases where the populations and goals differ.

In I/O psych and later education, the goal is often to contrast with academic ability, with some focus on interpersonal interactions (e.g., so-called bedside manner in medical students, whether an employee would work well with others). The population of interest is mostly young adults and older teens who have already surpassed some academic thresholds to reach this point in education or the workplace. Within that already narrowed population, then, educators and employers seek to differentiate the pool further to predict workplace effectiveness and/or select best candidates for specific types of positions. Most often, therefore, non-cognitive assessments are used as predictors, but the related skills may also be measured as an outcome of targeted training (e.g., bedside manner). In some parts of the literature, the terminology suggests assumptions of stability or inherent differences between individuals (e.g. "traits" or "attributes"), which aligns best with using assessments as predictors. When assessments are designed to measure outcomes, however, this indicates some expectation that non-cognitive skills are malleable. Assessments are designed differently when the construct of interest is assumed to be stable versus malleable, but these assumptions are not often made explicit when presenting assessment results.

Similar to I/O psych and later education, non-cognitive assessments in early education are designed to contrast with academic ability, and may also touch on interpersonal interactions (e.g., anti-bullying efforts). A major difference in early education, however, is the focus on children and adolescents, who have very different base capabilities for assessments. This necessarily changes the types of assessments that can be used, as we discuss further below. Another differentiation is the inclusion of all students, with a goal to support the development of these skills. Because research has shown non-cognitive factors as predictors of later success in education and employment, teachers, administrators, and parents are motivated to support these factors to benefit students' academic achievement, retention, continuation. In these younger grades, then, non-cognitive skills may be targeted with a specific curriculum and then measured as outcomes. Early research has also sought to establish typical and atypical developmental trajectories of these factors in the absence of curricular guidance.

2.2 Comparing Assessment Types: Examples from Self-control

To make concrete the types of challenges to assessment that must be addressed, we will use self-control as an example. Some of these challenges are general to questionnaires versus performance tasks (see Duckworth & Yeager, 2015, for further discussion of task types and the related threats to validity), but others are specific to underlying assumptions about the constructs of interest.

Questionnaires. One well-cited example comes from Moffitt et al. (2011) who argued that many important outcomes in adulthood are predicted by self-control during early childhood to adolescence. They followed 1000 participants in New Zealand from birth to 32 years of age and used childhood self-control as a predictor of later outcomes (four health dimensions, six wealth dimensions, and criminal convictions). They used a composite of self-control measures across age and observers, as described further below. The authors reported that the effects of self-control on later outcomes were dissociable from IQ (average scores on the Wechsler Intelligence Scale for Children—Revised, measured at 7, 9, and 11 years of age), socio-economic status (1–6 scale based on parent occupation), and so-called "adolescent snares" (e.g. early smoking, parenthood, school drop-out). We chose this study as an illustration because the findings have been influential on the literature, the measures are typical of these types of large-scale predictive studies, and it used a range of assessments at different ages in childhood.

The measures used by Moffitt et al. (2011), as described in their supplementary materials, were questionnaires developed as assessments of personality within social psychology or psychiatry. One implication of these disciplinary origins is an assumption that these are stable traits being measured, which influences the way the questionnaires were developed. Due to the wide age range covered, Moffitt et al. used different informants at different ages. The informants at ages 3 and 5 years were from trained observers who were part of the research team; they rated the child's lack of control (item content, see supplementary materials p. 2: "labile, low frustration tolerance, lack of reserve, resistance, restless, impulsive, requires attention, brief attention to task, lacks persistence in reaching goals"). At ages 5, 7, 9, and 11 years, parents and teachers reported on impulsive aggression ("flies of handle, fights"), hyperactivity ("runs and jumps about, cannot settle, has short attention span" at all ages; at 9 and 11 only, "'on the go' as if 'driven by a motor', difficulty sitting still". At 9 and 11 years, teachers and parents also rated lack of persistence ("fails to finish tasks, easily distracted, difficulty sticking to activity") and impulsivity ("acts before thinking, has difficulty awaiting turn, shifts excessively between activities"). At 11, adolescents self-rated hyperactivity ("fidgety, restless"), inattention ("difficulty paying attention, trouble sticking to a task"), and impulsivity ("difficulty waiting turn, talking while others are still talking").

There are a number of threats to validity and reliability in these types of methods (see Duckworth & Yeager, 2015, for a related overview and discussion of measurement issues), which are not specific to the Moffitt et al. (2011) study but are general to questionnaire assessments. Many techniques have been developed to address these threats, but none can be perfect, and across studies they may not be used consistently or details of implementation can be difficult to find (e.g., if specific items are not made available, general instructions before the questionnaire are not described). Furthermore, some of these techniques have been developed using narrow samples and may not be equipped to handle cultural differences (some of which we highlight below).

Source of Information. The first threat to validity and reliability of questionnaires we consider is the source. Moffitt et al. (2011) used all three of the types of informants who may be used in these assessments: (1) trained observers from the research team, (2) people in the subject's life who know them well (teachers and parents in Moffitt et al.; other studies have used coworkers or classmates), and (3) the subjects themselves. Each of these has potential concerns with framing, opportunity to observe, and bias, which we consider in turn here. Framing concerns who the reference group is for the subject, which can be defined at the outset of the questionnaire (e.g., students in the same school, children of the same age) for all informants. However, only with trained observers is there an opportunity to carefully control knowledge of the reference group—for example, if evaluating a subject relative to other children of the same age, how much the observer knows about typical children of that age can be evaluated and supported through training. Other informants' knowledge of the reference group could be measured within a questionnaire, but often researchers do not want to extend the length any more than is necessary, and these items are not central to the research questions of most studies. Proxies for knowledge can be designed with only a few items (e.g., for teachers, "How many years have you taught this grade?") but provide only a coarse metric.

Different informants will also have different opportunities to observe behaviors in question. Only subjects themselves have universal access to the subject's behavior, so any outside observer would have a more limited opportunity to observe particular behaviors. While the concern with framing is best handled by using a trained observer, this type of informant is the most limited in the opportunity to observe. This problem is exacerbated by the fact that behaviors are likely contextually-bound (discussed below), and therefore the level of self-control a child demonstrates will differ between school and home and an unfamiliar setting (as is often the case with trained observers), as well as between settings or contexts within the school and home. If the goal of assessing self-control is to predict its effect in particular settings, then it is sensible to limit the informants to the relevant setting. Some cases are not clear-cut, however, as in the case of homework for example. Parents and teachers may legitimately rate a child's ability to complete homework differently when observed at home (the process of completion) versus school (the outcome of completion) because the contextual considerations will be different in those settings: a child may finish an assignment at home, but fail to find it when relevant at school the next day. Which part of this behavior is most relevant to the assessment may not be clear.

Even with ample opportunity to observe and good knowledge of the reference group, any informant may be motivated, consciously or unconsciously, to respond more or less favorably than their true knowledge of the subject. When evaluating one's self, the subject may want to appear better to an examiner, or may be overly critical of their own behaviors. Parents and teachers may want to reflect better on themselves if they view a child's behavior as their responsibility to support; conversely, they may view the child as more problematic on these dimensions due to other, unrelated behavioral issues. All observers may hold stereotypes (positive or negative) about the social groups the subject belongs to (e.g., gender, race, socioeconomic status, cultural background) that could lead them to interpret behavior as more in line with their expectations than the behavior truly is, or to more readily identify behaviors that violate stereotypes. Because biases could lead to either over- or under-estimation of particular behaviors, there is not an easy universal solution to adjust reporting. Furthermore, bias may be 'contagious' between informants. For example, children may internalize descriptions from parents and teachers, either negatively or positively. If a child has been described as easily distracted (on one extreme) versus attentive and diligent (on the other extreme), by the time they are old enough to provide self-ratings, they have likely received substantial feedback that could influence their ratings more than their actual behavior.

Finally, informants will also be required to make their evaluation of the subject concrete or quantifiable, a process that may introduce additional error or bias. Questionnaires typically ask observers to rate on a Likert-type scale covering a range of similarity to the subject or frequency of a behavior. These types of ratings require an internal threshold in the observer for what would constitute "a lot" or "a little" in terms of similarity to the subject or frequency. This likely relates closely to the source's opportunity to observe both the subject and members of the reference group, but could also differ between individuals with similar experiences. For frequency, it is possible to adapt scales to be more concrete (e.g., "at least once a week", "less than once a month") but even these ratings will rely on subjective recall.

Context-Dependence. Across all sources, the potential context-dependence of behaviors can make questionnaires an ill-suited assessment. Some behaviors of interest may only occur in particularly constrained contexts, for example, a child who has a problem maintaining attention on a book they are reading may be better able to attend to the book being read to them (or vice versa). Children may become deeply engrossed in subjects they enjoy but have trouble sticking to a task in a subject they do not enjoy. Because questionnaire items tend to ask for generalizations or typical behaviors, it is unclear how one should respond to an item that differs across contexts. Items may be constructed to address different contexts, for example by specifying whether an activity is one the child likes or dislikes, but individual differences in contextual variation make it difficult to design items that will work equally well for all subjects.

Performance Tasks. As a contrasting example to questionnaires as assessments of self-control assessments, we consider a performance task: the delay of gratification task (Mischel & Ebbesen, 1970; Mischel, Ebbesen, & Raskoff Zeiss, 1972; Mischel, Shoda, & Rodriguez, 1989), colloquially referred to as the 'marshmallow task'. In this task, the participant is told that they can either have a small snack (e.g., one marshmallow) immediately or a larger snack (e.g., two marshmallows) if they wait. The snack is typically presented in front of the participant during the delay to maximize temptation, although there are a number of variations in the method including the personalization of the snacks offered to ensure that different participants will be similarly tempted (see Duckworth & Kern, 2011, for additional variations). The dependent measure of interest is how long the participant waits to receive the larger snack. This task has shown longitudinal predictive power from childhood to adolescence (e.g., Mischel, Shoda, & Peake, 1988; Shoda, Mischel, & Peake, 1990), and correspondence to concurrent parent ratings in early childhood and teacher ratings in middle childhood (Duckworth, Tsukayama, & Kirby, 2013). Mischel's work was notable for challenging the characterization of personality traits as leading to stable behavioral patterns (e.g., Mischel & Shoda, 1995), and he presented the marshmallow task as a way to study different contextual influences on children's performance. Thus, although this performance task originated from the same discipline as many questionnaires described above, the assumptions underlying the design and goals of the assessment differed.

Benefits of performance tasks over questionnaires include the direct assessment of participants' choices, the possibility for parametric manipulation (e.g., length of delay, difference between small and large snacks, even broader context as described below). This task has been adapted across age groups from 3 years to adults by modifying the nature and size of the choices and delay to be appropriate for the sample. There are, however, still significant threats to validity and reliability of this task. It is unclear how stable performance should be when used repeatedly, as having performed the task before may change how participants approach and interpret the task. Variance across individuals can be truncated by the experimenter-imposed delay limit; for example, in one study 59% of 10-year-olds waited the maximum 30 min for the larger delay, providing no differentiation among individuals in that portion of the sample (Duckworth et al., 2013).

Although this type of task has been used extensively with preschool-aged children, it is unclear whether they understand the tasks as the experimenter intends and/or trust that experimenter will follow through on the promised reward. Trust in the larger reward is essential for the task to work, yet it is not easily measured in young children performing the task. Kidd, Palmeri, and Aslin (2013) showed that children's trust can be modulated by the context of the task: if in an earlier interaction the experimenter failed to provide something that was promised, children were less likely to wait for the larger reward (see Ma, Chen, Xu, Lee, & Heyman, 2018; Michaelson & Munakata, 2016, for related findings with children; and Michaelson, de la Vega, Chatham, & Munakata, 2013, for similar findings with adults). As further evidence for the influence of context, Doebel and Munakata (2018) showed that children's beliefs about whether others in their peer group waited influenced how long they were willing to wait.

Implications of Assumptions Underlying Assessment Design. A general question applicable to all methods of assessment is what the expected stability should be over time. The expectation likely differs by theoretical perspective and/or discipline, as mentioned above: if self-control is considered a *trait* of an individual, one might expect it to change very little; if it is viewed as a *skill*, one might expect it to improve over time with practice. In the examples we used above, reliability was relatively low. Moffitt et al. (2011) reported only a moderate correlation (r = 0.30) between a composite of the childhood measures and a survey of the subjects as adults (as rated by themselves and a friend, partner or family member). Depending on one's expectations and perspective, this could be viewed positively—child self-control provides additional explanatory power over adult self-control, which assumes they are separable constructs—or negatively—indicating that self-control is not a stable trait or that our measures cannot reliably assess it. For the marshmallow task, we found only one report of test-retest reliability, which was r = 0.22 (Mischel et al., 1988). As noted above, it is unclear if performance tasks are undermined by repeated testing (a possibility raised by Mischel et al., 1988), as participants might have different expectations or find new strategies to approach the task, which therefore could change the construct(s) they measure.

These assessment examples in self-control raise the question of whether it is best to use one measure (where threats may be well-known and attempted to control) or multiple measures (each with different limitations that may have opposite effects), and how any measure should deal with context sensitivity. Regardless of which choice researchers make, it is important that we acknowledge the limitations of each type of measure and work to account for these when drawing conclusions about the underlying constructs. These types of concerns generalize to other competencies as well, beyond the specific measures discussed above and the construct of self-control.

3 Outlook

The goal of this paper was to consider how non-cognitive assessments are shaped by assumptions about the underlying constructs within and between disciplines. Comparing across disciplines, we considered how different purposes, goals, and populations shape these assumptions, in turn influencing the way assessments are designed and interpreted. We then used specific examples of assessments of one type of non-cognitive factor, self-control, to illustrate how different assumptions about this construct play out in questionnaires versus performance tasks. In this section, we also considered the threats to validity and reliability of these different assessments.

Moving forward, we recommend not only that these issues around measurement be brought to the forefront, but also that the theoretical implications surrounding them. For example, context sensitivity should be thought of as a feature of human behavior, not a noise to be eliminated. There are no 'pure' underlying abilities devoid of the influence of behavior, whether that behavior is measured in a performance task or generating a rating for a questionnaire. We should measure within the context we care most about, repeatedly and frequently, expecting variation and potentially even using that variation as a metric of interest (cf. van Geert & van Dijk, 2002).

Lastly, at the broadest level, the field would benefit from more specificity in our descriptions of the constructs we seek to measure. Although referring to 'non-cognitive' skills carries with it connotations that researchers across disciplines identify with, trying to measure at a level so general will sacrifice precision and specificity in predictive power. We are doing ourselves a disservice by lumping together such disparate skills as collaboration and study skills, for example; although these do have some underlying commonalities like self-regulation, the variation in how self-regulation influences the overt behaviors we can measure will make it difficult to pull those commonalities out. As our abilities to collect and analyze more extensive and different types of data increase, we must be mindful of the limitations of existing assessments and consider new ways to overcome them.

References

Doebel, S., & Munakata, Y. (2018). Group influences on engaging self-control: Children delay gratification and value it more when their in-group delays and their out-group doesn't. *Psychological Science, 29*(5), 738–748. https://doi.org/10.1177/0956797617747367.

Duckworth, A. L., Tsukayama, E., & Kirby, T. A. (2013). Is it really self-control? Examining the predictive power of the delay of gratification task. *Personality and Social Psychology Bulletin, 39*(7), 843–855. https://doi.org/10.1177/0146167213482589.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher, 44*(4), 237–251.

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality, 45*(3), 259–268. https://doi.org/10.1016/j.jrp.2011.02.004.

Easton, J. (2013). Using measurement as leverage between developmental research and educational practice. In *Center for Advanced Study of Teaching and Learning Meeting*. Charlottesville, VA. Retrieved from http://ies.ed.gov/director/pdf/Easton062013.pdf.

Kidd, C., Palmeri, H., & Aslin, R. N. (2013). Rational snacking: Young children's decision-making on the marshmallow task is moderated by beliefs about environmental reliability. *Cognition, 126*(1), 109–114. https://doi.org/10.1016/j.cognition.2012.08.004.

Kyllonen, P. C., Bertline, J., & Zu, J. (2018, April). *Measuring hard-to-measure (noncognitive) skills: Social, emotional, self-management, and beyond*. Presented at the Pre-conference training session at the meeting of the National Council on Measurement in Education, New York, NY.

Ma, F., Chen, B., Xu, F., Lee, K., & Heyman, G. D. (2018). Generalized trust predicts young children's willingness to delay gratification. *Journal of Experimental Child Psychology, 169,* 118–125. https://doi.org/10.1016/j.jecp.2017.12.015.

Michaelson, L. E., de la Vega, A., Chatham, C. H., & Munakata, Y. (2013). Delaying gratification depends on social trust. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00355.

Michaelson, L. E., & Munakata, Y. (2016). Trust matters: Seeing how an adult treats another person influences preschoolers' willingness to delay gratification. *Developmental Science, 19*(6), 1011–1019. https://doi.org/10.1111/desc.12388.

Mischel, W., Shoda, Y., & Rodriguez, M. (1989). Delay of gratification in children. *Science, 244*(4907), 933–938. https://doi.org/10.1126/science.2658056.

Mischel, W., & Ebbesen, E. B. (1970). Attention in delay of gratification. *Journal of Personality and Social Psychology, 16*(2), 329.

Mischel, W., Ebbesen, E. B., & Raskoff Zeiss, A. (1972). Cognitive and attentional mechanisms in delay of gratification. *Journal of Personality and Social Psychology, 21*(2), 204–218. https://doi.org/10.1037/h0032198.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*(2), 246–268. https://doi.org/10.1037/0033-295X.102.2.246.

Mischel, W., Shoda, Y., & Peake, P. K. (1988). The nature of adolescent competencies predicted by preschool delay of gratification. *Journal of Personality and Social Psychology, 54*(4), 687–696. https://doi.org/10.1037/0022-3514.54.4.687.

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., … Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. In *Proceedings of the National Academy of Sciences* (Vol. 108, Issue 7, pp. 2693–2698). https://doi.org/10.1073/pnas.1010076108.

Shoda, Y., Mischel, W., & Peake, P. K. (1990). Predicting adolescent cognitive and self-regulatory competencies from preschool delay of gratification: Identifying diagnostic conditions. *Developmental Psychology, 26*(6), 978–986. https://doi.org/10.1037/0012-1649.26.6.978.

van Geert, P., & van Dijk, M. (2002). Focus on variability: New tools to study intra-individual variability in developmental data. *Infant Behavior and Development, 25*(4), 340–374.

An Attribute-Specific Item Discrimination Index in Cognitive Diagnosis

Lihong Song and Wenyi Wang

Abstract There lacks an item quality index as a measure of item's correct classification rates of attributes. The purpose of this study is to propose an attribute-specific item discrimination index as a measure of correct classification rate of attributes based on a q-vector, item parameters, and the distribution of attribute patterns. First, an attribute-specific item discrimination index was introduced. Second, a heuristic method was presented using the new index for test construction. The first simulation results showed that the new index performed well in that their values matched closely with the simulated correct classification rates of attributes across different conditions. The second simulation study results showed that the heuristic method based on the sum of the attributes' indices yielded comparable performance to the famous CDI. The new index provides test developers with a useful tool to evaluate the quality of diagnostic items. It will be valuable to explore the applications and advantages of using the new index for developing an item selection algorithm or a termination rule in cognitive diagnostic computerized adaptive testing.

Keywords Cognitive diagnosis · Item discrimination index · Correct classification rate · Test construction · The deterministic inputs · Noisy "and" gate model

1 Introduction

The primary objective of cognitive diagnosis is to classify examinees into latent classes determined by vectors of binary attributes. Thus, the statistical quality of diagnosis test or items is most directly relevant to the classification accuracy. Since the estimation of attribute patterns no longer involves a continuous measure of ability, indices initially introduced by classical test theory (CTT) and item response theory

L. Song (✉) · W. Wang
Jiangxi Normal University, Nanchang, Jiangxi, People's Republic of China
e-mail: viviansong1981@163.com

W. Wang
e-mail: wenyiwang@jxnu.edu.cn

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_16

(IRT), such as reliability and information, do not apply directly to discrete latent space modeled by cognitive diagnostic models (Henson & Douglas, 2005).

Item discrimination indices for cognitive diagnosis provide an important tool for understanding the statistical quality of a diagnostic item and identifying "good" items (Rupp, Templin, & Henson, 2010). There are two basic sets of item discrimination index to measure discriminatory power of an item (Rupp et al., 2010). The first one is based on descriptive measures from CTT, such as the global item discrimination index. The second index is based on information measures from IRT, including the cognitive diagnosis index (CDI; Henson & Douglas, 2005), the attribute discrimination index (ADI; Henson, Roussos, Douglas, & He, 2008), the modified CDI and ADI (Kuo, Pai, & de la Torre, 2016). The modified CDI and ADI (Kuo et al., 2016) take into account attribute hierarchy and the ratio of test length to the number of attributes for designing cognitive diagnostic assessment. Moreover, the mutual information reliability (MIR) coefficient is proposed to evaluate the measurement quality of latent class or attribute pattern obtained from a mastery or diagnostic test (Chen, Liu, & Xu, 2018). The MIR is also a kind of information-based discrimination index because it is defined as $1 - SHE(\boldsymbol{\alpha}|\mathbf{X})/SHE(\boldsymbol{\alpha})$, where $SHE(\boldsymbol{\alpha}|\mathbf{X})/SHE(\boldsymbol{\alpha})$ refers to the conditional entropy of attribute pattern $\boldsymbol{\alpha}$ given item responses $\mathbf{X}$ and $SHE(\boldsymbol{\alpha})$ refers to the shannon entropy of attribute pattern $\boldsymbol{\alpha}$.

A global item discrimination index as a CTT-based index is defined as $d_j = P(X_j|\boldsymbol{\alpha}_u) - P(X_j|\boldsymbol{\alpha}_l)$, where $P(X_j|\boldsymbol{\alpha})$ is the probability of a correct response of item j with attributes pattern $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}_u = (1, 1, \ldots, 1)$, and $\boldsymbol{\alpha}_l = (0, 0, \ldots, 0)$. Let K be the number of attributes. The corresponding attribute-specific item discrimination index is defined as $d_{jk} = P(X_j|\boldsymbol{\alpha}_u) - P(X_j|\boldsymbol{\alpha}_l)$, where $\alpha_{uk} = 1$, $\alpha_{lk} = 0$, and $\alpha_{uk'} = \alpha_{lk'}$ for any $k' \neq k$, $k', k = 1, 2, \ldots, K$. For the deterministic inputs, noisy "and" gate model (called the DINA model; Haertel, 1989; Junker & Sijtsma, 2001) in Sect. 2.1, if s_j and g_j are slip and guessing parameters, then $d_j = (1 - \mathrm{s}_j) - \mathrm{g}_j$. If attribute k is measured by item j, then $d_{jk} = (1 - \mathrm{s}_j) - \mathrm{g}_j$, otherwise $d_{jk} = 0$.

Cognitive diagnostic information (CDI; Henson, & Douglas, 2005) as a global information-based item discrimination index which is based on Kullback-Leibler information (Chang & Ying, 1996) can be written as

$$CDI_j = \frac{1}{2^K(2^K - 1)} \sum\nolimits_{u \neq v} D_{juv}, \tag{1}$$

where,

$$\begin{aligned} D_{juv} &= KLI_j(P(X_j|\boldsymbol{\alpha}_u), P(X_j|\boldsymbol{\alpha}_v)) = E_{\boldsymbol{\alpha}}\left[ln\left[\frac{P(X_j|\boldsymbol{\alpha}_u)}{P(X_j|\boldsymbol{\alpha}_v)}\right]\right] \\ &= \sum\nolimits_{x=0}^{1} P(X_j = x|\boldsymbol{\alpha}_u) ln\left[\frac{P(X_j = x|\boldsymbol{\alpha}_u)}{P(X_j = x|\boldsymbol{\alpha}_v)}\right]. \end{aligned} \tag{2}$$

Kullback-Leibler information or, more appropriately, the Kullback-Leibler information for discrimination (Lehmann & Casella, 1998) is most commonly thought of as a measure of distance between any two probability distributions, f(x) and g(x). Chang and Ying (1996) suggested using Kullback-Leibler information instead of Fisher information as a more effective index for item selection in computerized adaptive tests based on unidimensional IRT models. For the DINA model, the CDI can be written as

$$CDI_j = \frac{1}{2^K(2^K-1)}\left(2^K - 2^{K-K_j}\right)2^{K-K_j}\left(1-\mathrm{s}_j-\mathrm{g}_j\right)$$
$$(\log((1-\mathrm{s}_j)/\mathrm{g}_j)) + \log((1-\mathrm{g}_j)/\mathrm{s}_j). \tag{3}$$

There are two attribute-specific item discrimination indices (Henson et al., 2008) as a kind of information-based index

$$d_{(A)jk} = \frac{1}{2^{K-1}}\left(\sum\nolimits_{\Omega_{k1}} D_{juv} + \sum\nolimits_{\Omega_{k0}} D_{juv}\right), \tag{4}$$

$$d_{(B)jk} = \sum\nolimits_{\Omega_{k1}} w_{k1} D_{juv} + \sum\nolimits_{\Omega_{k0}} w_{k0} D_{juv}, \tag{5}$$

where K is the number of attributes,

$$\Omega_{k1} = \left\{(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)|\alpha_{uk} = 1, \alpha_{vk} = 0,\ and\, \alpha_{uk'} = \alpha_{vk'}, \forall k' = k\right\}, \tag{6}$$

$$\Omega_{k0} = \{(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)|\alpha_{uk} = 0, \alpha_{vk} = 1,\ and\, \alpha_{uk'} = \alpha_{vk'}, \forall k' = k\}, \tag{7}$$

$$w_{k1} = P(\boldsymbol{\alpha}_u|\alpha_{uk} = 1), \tag{8}$$

and

$$w_{k0} = P(\boldsymbol{\alpha}_u|\alpha_{uk} = 0). \tag{9}$$

From all indices above, they are dependent on the characteristics of items and population, such as the q-vector, item parameters, and the distribution of attribute patterns. Existing studies have shown that there are strong relationships between the discrimination indices and correct classification rates (Henson et al., 2008) or classification error (Chen et al., 2018). However, there lacks an item quality index which can directly measure or estimate item's correct classification rates of attributes. The purpose of this study is to propose an attribute-specific item discrimination index based on a q-vector, item parameters, and the distribution of attribute patterns, as a classification-based index for measuring the statistical quality of a diagnostic item in terms of classification error.

The remainder of this chapter is structured as follows. Firstly, an attribute-specific item discrimination index is proposed to serve as a measure of correct classification rates for attributes at the item level. Secondly, a heuristic method is presented using the new index for test construction. Thirdly, simulation studies and results are elaborated.

Finally, the conclusions and discussion on future research directions are summarized in the final section.

2 Methodology

2.1 The DINA Model

The DINA model, a commonly used cognitive diagnostic model, is chosen here because it is a parsimonious and interpretable model that requires only two item parameters for each item regardless of the number of attributes being measured. The item response function for the DINA model is given by

$$\mathrm{P}_j(\boldsymbol{\alpha}_i) = P\left(X_{ij} = 1|\boldsymbol{\alpha}_i\right) = \left(1 - s_j\right)^{\eta_{ij}} g_j^{1-\eta_{ij}}, \tag{10}$$

where the deterministic latent response $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ki}^{q_{kj}}$ indicates whether or not examinee i possesses all of the attributes required by item j, 1 if mastered and 0 if unmastered. Let $\boldsymbol{\alpha}_i$ denote an attribute mastery pattern or a knowledge state from the universal set of knowledge states. Moreover, Q-matrix that specifies the item-attribute relationship is a $K \times J$ matrix, in which entry $q_{kj} = 1$ if attribute k is required for answering item j correctly and $q_{kj} = 0$ otherwise. The slip parameter s_j refers to the probability of an incorrect response to the item j when $\eta_{ij} = 1$, and the guessing parameter g_j represents the probability of a correct response to item j when $\eta_{ij} = 0$. X_{ij} is a binary random variable denoting the response of examinee i to item j. In addition, let $K_j = \sum_{k=1}^{K} q_{kj}$ be the total number of attributes measured by item j.

2.2 An Attribute-Specific Item Discrimination Index

If an individual's status on attribute k is classified, respectively, as $\widehat{\alpha}_{k1}$ and $\widehat{\alpha}_{k0}$ given a correct response $X_j = 1$ and an incorrect response $X_j = 0$ on item j. Let $P(\alpha_k = \widehat{\alpha}_{k1}|X_j = 1)$ and $P(\alpha_k = \widehat{\alpha}_{k0}|X_j = 0)$ represent the conditional probabilities of the two states of attribute k (i.e., $\widehat{\alpha}_{k1}$ or $\widehat{\alpha}_{k0}$) given item response X_j. Let $P\left(X_j = 1\right)$ and $P\left(X_j = 0\right)$ denote the marginal probabilities of a correct response and an incorrect response on item j, respectively. Since $P(\alpha_k = \widehat{\alpha}_{k1}|X_j = 1)$ and $P(\alpha_k = \widehat{\alpha}_{k0}|X_j = 0)$ can be viewed as an attribute-level classification accuracy index (Wang, Song, Chen, Meng, & Ding, 2015) given item response X_j. In view of the randomness of item responses, by taking an expectation of the attribute-level classification accuracy index with respect to item responses, thus an attribute-

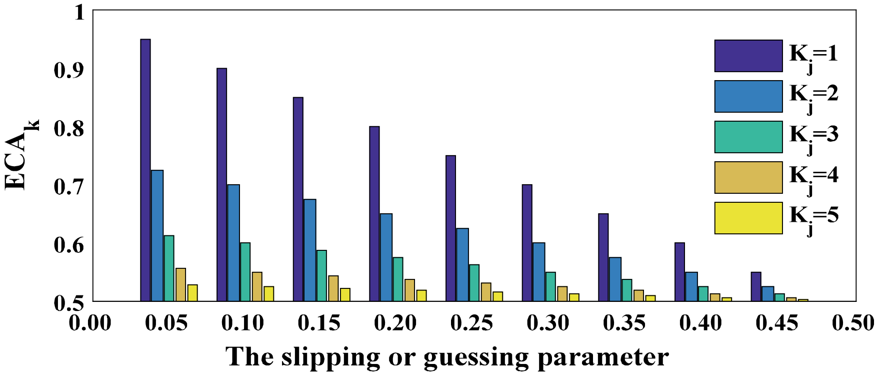


Fig. 1 The values of ECA_{jk}

specific item discrimination index or expected classification accuracy, denoted by ECA, can be written as

$$ECA_{jk} = P\big(\alpha_k = \hat{\alpha}_{k1} | X_j = 1\big) P\big(X_j = 1\big) + P\big(\alpha_k = \hat{\alpha}_{k0} | X_j = 0\big) P\big(X_j = 0\big). \tag{11}$$

Consider the expected a posteriori (EAP) estimate of attribute patterns (Huebner & Wang, 2011) under the DINA model. If the attributes are independent of each other and the attribute patterns follow a uniform distribution, then for $1 - s_j > g_j$, $q_{jk} = 1$, and $K_j = \sum_{k=1}^{K} q_{jk}$, the ECA for attribute k on item j is given by

$$ECA_{jk} = (1 - s_j - g_j)/2^{K_j} + 0.5. \tag{12}$$

Figure 1 shows the values of ECA_{jk} for items with different item parameters and the numbers of attributes required by items.

3 Simulation Study I

3.1 Study Design

The first simulation study was conducted to investigate whether the ECA index can accurately estimate the simulated or true values of correct classification rates of attributes. The true or simulated correct classification rate of an attribute was the proportion of times that each examinee was classified into the true attribute via the EAP given item responses on each item.

From Formulas (11) and (12), we know that the ECA index is dependent on three factors: (a) item parameters, (b) the number of attributes required by an item, and

Table 1 Conditions of simulation study

Factors	Levels
Model	The DINA model
The number of attributes	$K = 5$
The number of examinees	$N = 10{,}000$
The number of items	$J = 1000$
The Q-matrix	Each entry has a probability of 0.6 of being a 1
Item parameters	U(0.05, 0.25) and U(0.25, 0.45)
Distributions of attribute patterns	$\rho = 0$, 0.5, 0.75, and 0.95

Note For four distributions of attribute patterns ($\rho = 0$, 0.5, 0.75, and 0.95), the first distribution was a uniform distribution and the remaining three distributions were generated in the way of Chiu et al. (2009) and Henson et al. (2008)

(c) the distribution of attribute patterns. These three factors will be manipulated in this study. Table 1 shows a list of factors and their levels. The number of attributes measured by a test was fixed at $K = 5$.

Magnitude of item parameters. The size of item parameters was varied at two levels: low and high. Low level referred to relatively smaller guessing and slip parameters, which were randomly generated from a uniform distribution U(0.05, 0.25). High level, on the contrary, indicated relatively larger guessing and slip parameters, which were randomly generated from a uniform distribution U(0.25, 0.45) (Chen, Xin, Wang, & Chang, 2012).

The number of attributes required by an item. A total number of 1000 items (i.e., M = 1000) with two different distributions of item parameters was simulated to form two item banks. The elements of **Q**-matrix were generated item by item and attribute by attribute. To ensure that all the possible combinations of attributes or q-vectors occurred, the probability that an attribute was measured by an item was set to be 0.6 as in the studies of Cui, Gierl, and Chang (2012), and Wang et al. (2015).

The distribution of attribute patterns. Four distributions of attribute patterns were considered here. The first distribution was a uniform distribution, meaning that the attribute patterns were generated to take each of all 2^K possible patterns with equal probability. Considering the examinee's attribute-mastery probability and the relationship between the attributes, the remaining three distributions were generated in the following way (Chiu, Douglas, & Li, 2009; Henson & Douglas, 2005; Henson et al., 2008): first, continuous vectors of latent abilities were randomly drawn from a multivariate normal distribution $\boldsymbol{\theta} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\rho})$, where $\boldsymbol{\rho}$ represented a correlation matrix with equal off-diagonal elements. Three values of 0.5, 0.75 and 0.95 were assigned to the off-diagonal elements as in the study of Henson et al. (2008), respectively; second, the i-th individual's mastery for attribute k was given by

$$\alpha_{ik} = \begin{cases} 1 \; if \; \theta_{ik} \geq \phi^{-1}\left(\frac{k}{K+1}\right) \\ 0 \qquad\quad otherwise, \end{cases}. \tag{13}$$

where $k = 1, 2, \ldots, K$.

In this study, we simulated $N = 10{,}000$ examinees. An informative prior distribution estimated from the population of 10,000 examinees. The ECA index was computed from Formula (11) by employing different informative prior distributions corresponding to the four distributions.

3.2 Evaluation Criteria

The correct classification rates of attributes for the simulated data on each item was used as a baseline to evaluate the performance of the ECA. It should be noted that attribute pattern estimate $\widehat{\boldsymbol{\alpha}}_i$ was obtained via the EAP method with an informative prior distribution based on an item response of examinee i on a test item. Given the estimated attribute patterns $\widehat{\boldsymbol{\alpha}}_i$ and simulated $\boldsymbol{\alpha}_i$ $(i = 1, 2, \ldots, N)$, the correct classification rate for attribute k (Chen et al., 2012; Henson & Douglas, 2005; Kuo et al., 2016; Sun, Xin, Zhang, & de la Torre, 2013) was computed as follows

$$CCR_k = \frac{1}{N}\sum\nolimits_{i=1}^{N} I\left(\widehat{\alpha}_{ki}, \alpha_{ki}\right), \tag{14}$$

where $I(.)$ is an indicator function, having the value 1 for $\widehat{\alpha}_{ki} = \alpha_{ki}$ and the value 0 for $\widehat{\alpha}_{ki} \neq \alpha_{ki}$. Note that the CCR was considered as the true value of the ECA for attributes. With respect to the estimation precision, the overall recovery was evaluated by bias (BIAS), absolute bias (ABS), and root mean squared error (RMSE).

3.3 Results

Table 2 shows the estimation precision across all simulation conditions. On the whole, the BIAS was between −0.0248 and 0.0062; the ABS was between 0.0027 and 0.0333; the RMSE was between 0.0034 and 0.0390, indicating that all correct classification rates were recovered accurately by the ECA index.

As illustrated in the top left panel of Fig. 2, it was observed that the distribution of attribute patterns did not cause a significant impact on the estimation error. The impact of the number of attributes required by an item on the estimation precision is presented in the top right panel of Fig. 2. The estimation error on items with a higher number of attributes required by an item was slightly larger than that of items with a smaller number of attributes required by an item. This might be because the higher number of attributes, the more parameters needed to be estimated. The bottom panel of Fig. 2 shows the impact of item parameters on the estimation precision. Item parameters had a little impact on the accuracy of the proposed index.

Table 2 The precision of the ECA index across all simulation conditions

Correlation	Item parameters	K_{j}	BIAS	ABS	RMSE
ρ = 0	U(0.05, 0.25)	1	0.0006	0.0028	0.0035
		2	−0.0153	0.0154	0.0168
		3	−0.0248	0.0248	0.0268
		4	−0.0207	0.0208	0.0233
		5	−0.0143	0.0144	0.0167
	U(0.25, 0.45)	1	0.0007	0.0037	0.0045
		2	−0.0024	0.0044	0.0056
		3	−0.0032	0.0053	0.0066
		4	−0.0024	0.0051	0.0062
		5	0.0001	0.0045	0.0054
ρ = 0.5	U(0.05, 0.25)	1	−0.0001	0.0028	0.0034
		2	−0.0002	0.0188	0.0230
		3	−0.0008	0.0258	0.0295
		4	0.0025	0.0282	0.0323
		5	0.0047	0.0288	0.0329
	U(0.25, 0.45)	1	−0.0007	0.0028	0.0034
		2	0.0013	0.0071	0.0091
		3	0.0015	0.0076	0.0098
		4	0.0020	0.0072	0.0093
		5	0.0019	0.0060	0.0075
ρ = 0.75	U(0.05, 0.25)	1	0.0010	0.0030	0.0038
		2	0.0018	0.0185	0.0242
		3	−0.0005	0.0265	0.0314
		4	0.0015	0.0310	0.0363
		5	0.0032	0.0333	0.0390
	U(0.25, 0.45)	1	0.0011	0.0038	0.0048
		2	0.0026	0.0099	0.0119
		3	0.0031	0.0114	0.0140
		4	0.0038	0.0121	0.0148
		5	0.0040	0.0118	0.0144
ρ = 0.95	U(0.05, 0.25)	1	−0.0003	0.0027	0.0034
		2	0.0031	0.0148	0.0210
		3	0.0014	0.0213	0.0263
		4	0.0026	0.0257	0.0312
		5	0.0029	0.0285	0.0342

(continued)

Table 2 (continued)

Correlation	Item parameters	K_j	BIAS	ABS	RMSE
	U(0.25, 0.45)	1	0.0023	0.0033	0.0041
		2	0.0041	0.0064	0.0088
		3	0.0049	0.0077	0.0107
		4	0.0058	0.0085	0.0119
		5	0.0062	0.0085	0.0121
		M	−0.0004	0.0131	0.0158

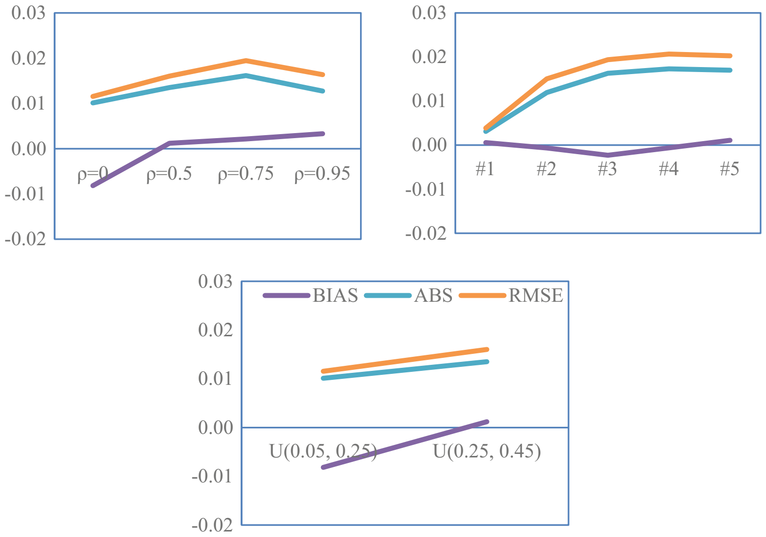


Fig. 2 The impact of correlations, the number of attributes and item parameters on the estimation precision

4 Simulation Study II

4.1 Study Design

The second study was conducted to investigate the performance of the ECA index for test construction, which was compared with the CDI (Henson & Douglas, 2005). Table 3 shows a list of factors and their levels. All data were simulated in the same way employed in the first study. Under an item bank with 300 items, a heuristic algorithm based on the ECA index for test construction used the following steps:

Table 3 Conditions of simulation study

Factors	Levels
Model	The DINA model
The number of attributes	$K = 4$ and 8
The number of examinees	$N = 10{,}000$
The number of items	$J = 300$
The Q-matrix	Randomly selecting from all possible attribute patterns with replacement
Item parameters	U(0.05, 0.40)
Distributions of attribute patterns	$\rho = 0$ and 0.5

Step 1. Select the first item with the largest ECA_j, where $\mathrm{ECA}_j = \sum_{k=1}^{K} \mathrm{ECA}_{jk}$.

Step 2. Select the next item such that ECA_j is the maximum of all remaining items in the item bank.

Step 3 Repeat step 2 until the desired test length (i.e. 40 or 80) is achieved.

4.2 Evaluation Criteria

Two hundred data sets of item responses were generated and analyzed for each condition. For each data set, attribute pattern $\widehat{\boldsymbol{\alpha}}_i$ was obtained via the EAP method using the prior distribution estimated from the population of 10,000 examinees. Given the estimated attribute patterns $\widehat{\boldsymbol{\alpha}}_i$ and simulated $\boldsymbol{\alpha}_i$ $(i = 1, 2, \ldots, N)$, the correct classification rate for attribute patterns (Chen et al., 2012; Henson & Douglas, 2005; Kuo et al., 2016; Sun et al., 2013) was computed as follows

$$CCR = \frac{1}{N} \sum\nolimits_{i=1}^{N} I\left(\widehat{\boldsymbol{\alpha}}_i, \boldsymbol{\alpha}_i\right), \tag{15}$$

where $I(.)$ is an indicator function, having the value 1 for $\widehat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i$ and the value 0 for $\widehat{\boldsymbol{\alpha}}_i \neq \boldsymbol{\alpha}_i$. The average of CCR across the 200 replications is computed for each condition.

4.3 Results

Figures 3 and 4 show the average correct classification rates of attribute patterns across the replications for two test construction methods under various conditions. The heuristic method based on the sum of the ECAs of attributes yielded comparable performance to the famous CDI across various test lengths regardless of correlations.

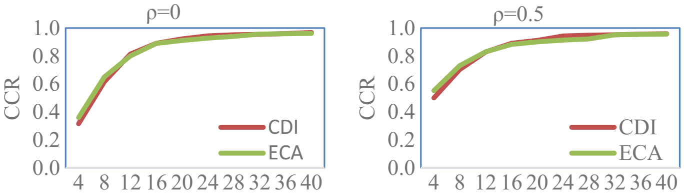


Fig. 3 Correct classification rates of attribute patterns for two test construction methods ($K = 4$)

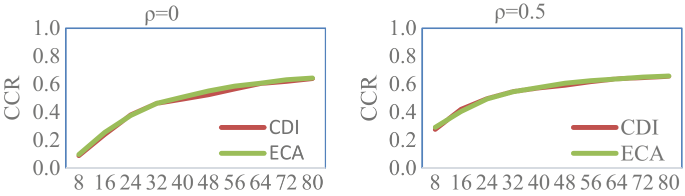


Fig. 4 Correct classification rates of attribute patterns for two test construction methods ($K = 8$)

5 Discussion

Theoretical results shows that the proposed attribute-specific item discrimination index, which is based on a q-vector, item parameters, and the distribution of attributes, can be viewed as a measure of correct classification rates of attributes. The first simulation study was conducted to evaluate the performance of the ECA under the DINA model. Several factors were manipulated for five independent attributes in this study. Results showed that the new index performed well in that their values matched closely with the simulated correct classification rates of attributes across different simulation conditions. The second simulation study was conducted to examine the effectiveness of the heuristic method for test construction. The test length was fixed to 40 or 80 and simulation conditions are similar as used in the first study. Results showed that the heuristic method based on the sum of the ECAs yielded comparable performance to the famous CDI.

These indices can provide test developers with a useful tool to evaluate the quality of the diagnostic items. The attribute-specific item discrimination index will provide researchers and practitioners a way to select the most appropriate item and test that they want to measure with greater accuracy. With these indices, one can get reliable statistical evidence of the quality of a single item and obtain valuable data for test construction in cognitive diagnostic assessment as well. Because the ECA can be regarded as a measure of correct classification rate of attributes on test items (Wang, Song, & Ding, 2018). Wang, Song, Chen, and Ding (2019) have proposed a method for making the prediction of the test-level correct classification rates of attribute patterns based on the ECAs of test items. Since computerized adaptive testing with cognitive diagnosis (CD-CAT) can help with classroom assessment and facilitate individualized learning (Chang, 2015), it is interesting to study the application of the designed index in CD-CAT for developing an item selection algorithms or a termination rule in future.

There were several limitations to this research. First, the current study focused on the DINA model only. In the future, the proposed index should be applied to general families of cognitive diagnostic models (CDMs) such as the generalized DINA model (de la Torre, 2011), the log-linear CDM (Henson, Templin, & Willse, 2009), the general diagnostic model (von Davier, 2005), and the polytomous CDMs (Xia, Mao, & Yang, 2018). Finally, the existing methods for test construction did not incorporate the informative prior distribution of attribute patterns into test construction. When the informative prior distribution is known, then Bayesian methods, which make use of more information, are preferred due to their better performance (Theodoridis & Koutroumbas, 2009). For example, the fully Bayes approach is most accurate, provided that the informative prior information is reliable (de la Torre, Hong, & Deng, 2010), and employing informative prior in item selection can significantly improve classification accuracy in CD-CAT (Cheng, 2009). Therefore, it is worth exploring ways of utilizing prior knowledge to develop new methods of test construction for further improving the classification accuracy of a diagnostic test in future.

Acknowledgments This research was supported by the Key Project of National Education Science "Twelfth Five Year Plan" of Ministry of Education of China (Grant No. DHA150285). The authors would like to thank the editor Steve Culpepper for reviewing an earlier version of this work.

References

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika, 80*(1), 1–20.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*(3), 213–229.

Chen, P., Xin, T., Wang, C., & Chang, H.-H. (2012). On-line calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika, 77*(2), 201–222.

Chen, Y., Liu, Y., & Xu, S. (2018). Mutual information reliability for latent class analysis. *Applied Psychological Measurement, 42*(6), 460–477.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD–CAT. *Psychometrika, 74*(4), 619–632.

Chiu, C.-Y., Douglas, J. A., & Li, X.-D. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*(4), 633–665.

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

de la Torre, J., Hong, Y., & Deng, W. L. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement, 47*(2), 227–249.

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.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement, 29*(4), 262–277.

Henson, R., Roussos, L., Douglas, J., & He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement, 32*(4), 275–288.

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.

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement, 71*(2), 407–419.

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.

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*(5), 315–330.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). New York: Springer.

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

Sun, J., Xin, T., Zhang, S., & de la Torre, J. (2013). A polytomous extension of the generalized distance discriminating method. *Applied Psychological Measurement, 37*(7), 503–521.

Theodoridis, S., & Koutroumbas, K. (2009). *Pattern recognition* (4th ed.). London: Elsevier.

von Davier, M. (2005). *A general diagnostic model applied to language testing data (ETS RR-05–16)*. Princeton, NJ: Educational Testing Service.

Wang, W. Y., Song, L. H., Chen, P., Meng, Y. R., & Ding, S. L. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement, 52*(4), 457–476.

Wang, W. Y., Song, L. H., & Ding, S. L. (2018). An item discrimination index and its application in cognitive diagnostic assessment on a classification oriented view. *Journal of Psychological Science, 41*(2), 475–483.

Wang, W. Y., Song, L. H., Chen, P., & Ding, S. L. (2019). An item-level expected classification accuracy and its applications in cognitive diagnostic assessment. *Journal of Educational Measurement, 56*(1), 51–75.

Xia, M. L., Mao, X. Z., & Yang, R. (2018). Cognitive diagnosis models under polytomous attributes and polytomous item. *Journal of Jiangxi Normal University (Natural Science), 42*(2), 134–138.

Assessing the Dimensionality of the Latent Attribute Space in Cognitive Diagnosis Through Testing for Conditional Independence

Youn Seon Lim and Fritz Drasgow

Abstract Cognitive diagnosis seeks to assess an examinee's mastery of a set of cognitive skills called (latent) attributes. The entire set of attributes characterizing a particular ability domain is often referred to as the latent attribute space. The correct specification of the latent attribute space is essential in cognitive diagnosis because misspecifications of the latent attribute space result in inaccurate parameter estimates, and ultimately, in the incorrect assessment of examinees' ability. Misspecifications of the latent attribute space typically lead to violations of conditional independence. In this article, the Mantel-Haenszel statistic (Lim & Drasgow in J Classif, 2019) is implemented to detect possible misspecifications of the latent attribute space by checking for conditional independence of the items of a test with parametric cognitive diagnosis models. The performance of the Mantel-Haenszel statistic is evaluated in simulation studies based on its Type-I-error rate and power.

Keywords Cognitive diagnosis model · Dimensionality · Mantel-haenszel statistic

1 Introduction

Cognitive diagnosis models (CDMs) try to account for the dependence among observations by latent dimensions that are related to the mastery or possession of cognitive skills, or "attributes" required for a correct response to an item. These models have received considerable attention in educational research because tests based on CDMs

Y. S. Lim (✉)
Department of Science Education,
Donald and Barbara Zucker School of Medicine at Hofstra/Northwell,
500 Hofstra University, Hempstead, NY 11549, USA
e-mail: younseon.lim@gmail.com; Younseon.Lim@hofstra.edu

F. Drasgow
School of Labor & Employment Relations, Department of Psychology, 504 E. Armory Avenue,
Champaign, IL 61820, USA

University of Illinois at Urbana-Champaign, 603 E. Daniel Street, Champaign, IL 61820, USA
e-mail: fdrasgow@illinois.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_17

promise to provide more diagnostic information about an examinee's ability than tests that are based on Item Response Theory (IRT) (Rupp et al., 2010). Specifically, whereas IRT defines ability as a unidimensional continuous construct, CDMs describe ability as a composite of K discrete, binary latent skill variables called attributes that define 2^K distinct classes of proficiency.

Like with other measurement models in assessment, the validity of a CDM depends on whether the latent attributes characterizing each proficiency class entirely determine an examinee's test performance, so that item responses can be assumed to be independent after controlling for the effect of the attributes. (This property of conditional independence is often called "local independence" in the IRT literature.) As Lord and Novick (1968) pointed out, the misspecification of the latent ability space underlying a test usually leads to violations of the conditional independence assumption that, in turn, result in inaccurate estimates of the model parameters and, ultimately, incorrect assessments of examinees' ability. For cognitive diagnosis, the assumption of conditional independence is equivalent to the assumption that the K attributes span the complete latent space. More to the point, violations of conditional independence are likely to occur if the latent attribute space has been misspecified in either including too few or too many latent attributes in the model.

Within the context of IRT models, various methods have been proposed for examining the dimensionality of the latent ability space underlying a test through checking for possible violations of conditional independence. Stout (1987), for example, developed DIMTEST, a nonparametric procedure for establishing unidimensionality of the test items through testing for conditional independence. Another instance is Rosenbaum's (1984) use of the Mantel-Haenszel statistic for assessing the unidimensionality of dichotomous items.

Lim and Drasgow (2019) proposed a nonparametric procedure for detecting misspecifications of the latent attribute space in cognitive diagnosis, which relies on the Mantel-Haenszel statistic to check for violations of conditional independence in the context of nonparametric cognitive diagnosis method approaches. This study extends the study of Lim and Drasgow (2019) by using the proposed statistic with parametric cognitive models for the estimation of proficiency classes.

2 The Mantel-Haenszel Test

Lim and Drasgow (2019) propose to use the Mantel-Haenszel (MH) chi-square statistic to test for the (conditional) independence of two dichotomous variables j and j' by forming the 2-by-2 contingency tables in conditioning on the levels of the stratification variable C. In their study, the stratification variable C is defined in terms of the latent attribute vector $\boldsymbol{\alpha}_c = (\alpha_{c1}, \alpha_{c2}, ..., \alpha_{cK})'$, for $c = 1, 2, ..., 2^K$; that is, the different strata of C are formed by the 2^K proficiency classes.

Let $\{i_{j,j'c}\}$ denote the frequencies of examinees in the $2 \times 2 \times C$ contingency table. The marginal frequencies are the row totals $\{i_{1+c}\}$ and the column totals $\{i_{+1c}\}$, and i_{++c} represents the total sample size in the cth stratum. Then, the MH statistic is defined as

$$\mathrm{MH}\chi^2 = \frac{[\sum_c (i_{11c} - \sum_c E(i_{11c})]^2}{\sum_c \mathrm{var}(i_{11c})}, \tag{1}$$

where $E(i_{11c}) = i_{1+c} i_{+1c} / i_{++c}$ and $\mathrm{var}(i_{11c}) = i_{0+c} i_{1+c} i_{+0c} i_{+1c} / i_{++c}^2 (i_{++c} - 1)$. The stratum having minimum total sample size i_{++c} equal or larger than 1 is included. Under the null hypothesis of conditional independence of the items j and j', for cognitive diagnosis models, the MH statistic has approximately a chi-square distribution with degrees of freedom equal to 1 if examinee's true latent attribute vectors are used as the levels of the stratification variable C. Assume that the odds ratio between j and j' is constant across all strata. Then the null hypothesis of independence is equivalent to an odds ratio of one

$$\text{Odds Ratio}_{\mathrm{MH}j,j'} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{or}_{j,j'c}, \tag{2}$$

where $\mathrm{or}_{j,j'c} = (i_{11c} i_{00c}) / (i_{10c} i_{01c})$.

3 Simulation Studies

The finite test-length and sample-size properties of $\mathrm{MH}\chi^2$ have been investigated in simulation studies. For each condition, item response data of sample sizes $I = 500$, or 2000 were drawn from a discretized multivariate normal distribution $MVN(0_K, \sum)$, where the covariance matrix $\sum$ has unit variance and common correlation $\rho = 0.3$ or 0.6. The K-dimensional continuous vectors $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, ..., \theta_{iK})'$ were dichotomized according to

$$\alpha_{ik} = \begin{cases} 1, & \text{if } \theta_{ik} \geq \Phi^{-1} \frac{k}{K+1}; \\ 0, & \text{otherwise} \end{cases}$$

Test lengths $J = 20$ or 40 were studied with attribute vectors of length $K = 3$ or 5. The correctly specified Q-matrix for $J = 20$ is presented in Table 1 (Attributes with ⋆ were used for Q-matrix ($K = 3$); attributes with ⋆⋆ for Items 4 and 5). The Q-matrix for $J = 40$ was obtained by duplicating this matrix two times.

Data were generated from three different models: the DINA model, the additive-cognitive diagnosis model (A-CDM), and a saturated model (i.e., the generalized-DINA (G-DINA) model). For the DINA model, item parameters were drawn from Uniform (0, 0.3). For the A-CDM and the saturated model, like Chen et al. (2013), the parameters were restricted as $P(\alpha_{ij}^{\star})_{\min} = 0.10$ and $P(\alpha_{ij}^{\star})_{\max} = 0.90$, where $\alpha_{ij}^{\star}$ was the reduced attribute vector whose components are the required attributes for the $j - th$ item (see de la Torre, 2011, more details). The R was used for the estimation in this study (e.g., Robitzsch, Kiefer, George, & Uenlue, 2015) in which model parameter estimation was performed by maximization of the marginal likelihood.

Table 1 Correctly specified Q (K = 5)

Item	$k_1^\star$	$k_2^\star$	$k_3^\star$	k_4	k_5	Item	$k_1^\star$	$k_2^\star$	$k_3^\star$	k_4	k_5
1	1	0	0	0	0	11	1	1	0	0	0
2	0	1	0	0	0	12	1	1	0	0	1
3	0	0	1	0	0	13	1	0	0	1	1
4	$0^{\star\star}$	0	0	$1^{\star\star}$	$0^{\star\star}$	14	0	1	0	1	1
5	$0^{\star\star}$	0	0	$0^{\star\star}$	$1^{\star\star}$	15	0	0	1	1	0
6	1	1	0	1	1	16	1	0	0	1	0
7	1	0	0	1	0	17	0	1	0	0	1
8	0	1	0	1	0	18	0	0	1	0	0
9	0	0	1	1	0	19	1	0	0	1	0
10	0	1	0	1	0	20	1	0	0	1	1

For each condition, a set of item response vectors was simulated for 100 replications. The proposed MH statistic, Chi-squared statistic $x_{jj'}$ (Chen and Thissen, 1997), absolute deviations of observed and predicted corrections $r_{jj'}$ (Chen et al. 2013), and their corresponding p-values were computed for all $(J \times (J-1))/2$ item-pairs in an individual replication.

4 Results

Across 100 trials for each condition, the proportion of times the p-value of each item-pair was smaller than the significance level 0.05 was recorded and is summarized in the tables shown below.

Type I Error Study In this simulation study, the correctly specified Q-matrices (K = 5, or K = 3) were used to fit the data to examine type I error rates. Table 2 shows that most type I error rates of the three different statistics were around the nominal significance level 0.05. The Chi-squared test statistic $x_{jj'}$ was conservative, with type I error rates below 0.024. The MH statistic got consistent under all conditions when item J = 40, confirming the asymptotic consistency. In the condition of K= 5, J = 20, and I = 2000, the type I error rates of the MH test slightly increased over the nominal rate in the A-CDM and the saturated model for the difficulty of correct classification.

Power Study: 20% misspecified Q-matrix For each replication, 20% of q_{jk} entries of the correctly specified Q-matrices (K = 5, or K = 3) were randomly misspecified. It is over-specification when q-entries of 0 are incorrectly coded as 1, and it is underspecification when q-entries of 1 are incorrectly coded as 0. Table 3 shows that the average rejection rates of all $J \times (J-1) \times 1/2$ item pairs result in relatively low in the MH test (i.e., 0.310 or below in Non-Parametric Model, 373 or below in

Table 2 Type I error study

		$J = 20$						$J = 40$					
		α with $\rho = 0.3$			α with $\rho = 0.6$			α with $\rho = 0.3$			α with $\rho = 0.6$		
K	I	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$
DINA model													
3	500	0.042	0.019	0.044	0.045	0.014	0.033	0.048	0.017	0.053	0.042	0.020	0.053
	2000	0.046	0.023	0.052	0.045	0.015	0.033	0.049	0.019	0.052	0.048	0.019	0.045
5	500	0.040	0.016	0.042	0.040	0.010	0.035	0.047	0.017	0.050	0.049	0.011	0.039
	2000	0.039	0.011	0.033	0.053	0.008	0.027	0.046	0.024	0.059	0.049	0.014	0.043
A-CDM													
3	500	0.036	0.009	0.029	0.031	0.009	0.026	0.039	0.011	0.030	0.036	0.011	0.028
	2000	0.048	0.013	0.030	0.049	0.010	0.026	0.048	0.010	0.029	0.047	0.010	0.026
5	500	0.021	0.020	0.038	0.024	0.015	0.033	0.039	0.018	0.046	0.034	0.014	0.039
	2000	0.073	0.016	0.036	0.065	0.011	0.031	0.040	0.019	0.047	0.043	0.014	0.039
Saturated model													
3	500	0.034	0.010	0.025	0.033	0.009	0.026	0.040	0.010	0.029	0.035	0.011	0.028
	2000	0.047	0.010	0.028	0.045	0.010	0.025	0.046	0.010	0.029	0.047	0.009	0.026
5	500	0.024	0.014	0.036	0.020	0.013	0.031	0.041	0.018	0.047	0.035	0.015	0.041
	2000	0.067	0.012	0.034	0.057	0.010	0.030	0.041	0.018	0.045	0.043	0.014	0.039

Table 3 Power study: 20% misspecified Q.

		$J = 20$						$J = 40$					
		α with $\rho = 0.3$			α with $\rho = 0.6$			α with $\rho = 0.3$			α with $\rho = 0.6$		
K	I	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$
DINA model													
3	500	0.132	0.250	0.305	0.117	0.210	0.271	0.111	0.266	0.322	0.100	0.238	0.296
	2000	0.235	0.400	0.449	0.209	0.393	0.472	0.143	0.397	0.438	0.134	0.379	0.430
5	500	0.161	0.199	0.275	0.143	0.168	0.239	0.111	0.257	0.330	0.124	0.242	0.326
	2000	0.373	0.465	0.536	0.294	0.360	0.448	0.105	0.186	0.263	0.207	0.394	0.470
A-CDM													
3	500	0.164	0.352	0.390	0.165	0.362	0.399	0.105	0.341	0.373	0.113	0.366	0.396
	2000	0.241	0.478	0.510	0.258	0.456	0.491	0.141	0.423	0.454	0.156	0.434	0.463
5	500	0.121	0.149	0.182	0.074	0.131	0.181	0.047	0.200	0.249	0.049	0.174	0.232
	2000	0.149	0.196	0.230	0.169	0.270	0.338	0.106	0.347	0.389	0.124	0.347	0.387
Saturated model													
3	500	0.118	0.342	0.388	0.110	0.337	0.372	0.101	0.336	0.379	0.090	0.348	0.384
	2000	0.206	0.470	0.493	0.270	0.578	0.614	0.131	0.441	0.473	0.130	0 .459	0.490
5	500	0.062	0.139	0.189	0.056	0.110	0.164	0.048	0.187	0.238	0.041	0.162	0.216
	2000	0.152	0.285	0.343	0.139	0.230	0.287	0.102	0.301	0.349	0.081	0.279	0.333

DINA model, 0.258 or below in A-CDM, 0.270 or below in saturated model). When $K = 5$, and $I = 500$, the power rates appear to be low (i.e., 0.074 or below) in the A-CDM, and the saturated model. They are rather complex models. It is very likely for small sample size to increase the difficulty of accurate model estimation.

Power Study: Over-specified *Q*-matrix For each replication, a data set was generated with the Q-matrix ($K = 3$) that is embedded as a subset of the Q-matrix ($K = 5$) in Table 1. The data was fitted with the Q-matrix ($K = 5$) to over-specify the correctly specified Q-matrix ($K = 3$). A dimension (total 9 items) or two dimensions (total 4 items) were over-specified. The results were consistent with what Chen et al. (2013) found.

As Table 4 shows, all statistics were insensitive to over-specified Q-matrices when the true models were the saturated model or the A-CDM. The average power rates of the item pairs where both items were over-specified in the same dimension were Non-Parametric Model = 0.074, MH = 0.052, $x_{jj'} = 0.181$, and $r_{jj'} = 0.220$, and those of the item pairs where either item was over-specified were MH = 0.058, $x_{jj'} = 0.104$, and $r_{jj'} = 0.137$ when the true model was the DINA model. If more attributes are included in the Q-matrix than required, as Rupp et al. (2010) indicated, conditional independence may still be preserved, because true attribute vector may be embedded in subcomponents of the modeled vector, resulting in a model that is too complex but preserves conditional independence. This finding implies that unlike the other statistics, the MH statistic is inappropriate to be used for the detection of the over-specified Q-matrices when the true model is the DINA model.

Power Study: Under-specified *Q*-matrix A data set was generated with the Q-matrix ($K = 5$) in Table 1. The data was fitted with the embedded Q-matrix ($K = 3$) in each replication. A dimension (total 9 items) or two dimensions (total 4 items) were under-specified. The average power rates of the item pairs where both items were under-specified in the same dimension were MH = 0.572, $x_{jj'} = 0.669$, and $r_{jj'} = 0.735$, with power relatively consistent across all conditions as shown in Table 5. The average rejection rates across item pairs where either item was under-specified were MH = 0.124, $x_{jj'} = 0.144$, and $r_{jj'} = 0.201$. The power rates slightly increased when $J = 40$, $I = 2000$, or the true model is the A-CDM in all statistics. Taking this finding into account, like the other statistics, the MH test is sensitive to Q-underspecification and has high power in all conditions.

Power Study: Model Misspecification In this simulation study, a correctly specified Q-matrix ($K = 3$ or 5) was used, but with a misspecified cognitive diagnosis models. As Chen et al. (2013) indicated, no statistics detected the model misspecification in all conditions when the fitted model was the saturated model, and the true models were the DINA model and the A-CDM (i.e., 0.052 or below for MH, 0.024 or below for $x_{jj'}$, and 0.059 or below for $r_{jj'}$). Due to limited space, the output is not included. The results in Table 6 show that the rejection rates of the MH statistic were low (i.e.,

Table 4 Power study: Over-specified Q with true $K = 5$.

		$J = 20$						$J = 40$					
		α with $\rho = 0.3$			α with $\rho = 0.6$			α with $\rho = 0.3$			α with $\rho = 0.6$		
I	S	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$
DINA model													
500	O	0.038	0.167	0.218	0.043	0.150	0.191	0.048	0.145	0.179	0.048	0.137	0.168
	E	0.059	0.095	0.128	0.046	0.080	0.118	0.051	0.086	0.114	0.049	0.086	0.115
2000	O	0.073	0.271	0.311	0.063	0.249	0.295	0.050	0.181	0.221	0.048	0.150	0.177
	E	0.075	0.139	0.175	0.079	0.156	0.194	0.048	0.100	0.134	0.056	0.094	0.119
A-CDM													
500	O	0.031	0.008	0.021	0.032	0.009	0.018	0.037	0.006	0.021	0.033	0.003	0.016
	E	0.047	0.010	0.024	0.040	0.008	0.024	0.040	0.006	0.023	0.039	0.006	0.024
2000	O	0.039	0.012	0.020	0.038	0.019	0.025	0.045	0.004	0.018	0.044	0.003	0.015
	E	0.040	0.011	0.024	0.043	0.013	0.025	0.046	0.006	0.023	0.046	0.006	0.022
Saturated model													
500	O	0.035	0.007	0.020	0.038	0.004	0.016	0.038	0.007	0.022	0.033	0.003	0.017
	E	0.052	0.006	0.020	0.049	0.004	0.018	0.040	0.007	0.027	0.039	0.005	0.023
2000	O	0.041	0.012	0.019	0.041	0.085	0.096	0.044	0.006	0.021	0.042	0.005	0.017
	E	0.042	0.018	0.020	0.041	0.045	0.058	0.042	0.007	0.024	0.046	0.007	0.023

Table 5 Power study: under-specified Q with true K=6

		$J = 20$						$J = 40$					
		α with $\rho = 0.3$			α with $\rho = 0.6$			α with $\rho = 0.3$			α with $\rho = 0.6$		
I	S	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$
DINA model													
500	U	0.390	0.462	0.541	0.672	0.595	0.725	0.488	0.550	0.608	0.319	0.415	0.499
	E	0.093	0.088	0.149	0.150	0.064	0.136	0.145	0.130	0.194	0.094	0.129	0.194
2000	U	0.532	0.608	0.648	0.489	0.599	0.672	0.538	0.624	0.658	0.780	0.845	0.900
	E	0.210	0.245	0.314	0.106	0.171	0.225	0.290	0.285	0.359	0.189	0.180	0.259
A-CDM													
500	U	0.464	0.553	0.637	0.453	0.581	0.675	0.551	0.642	0.722	0.535	0.682	0.766
	E	0.093	0.061	0.100	0.064	0.081	0.124	0.051	0.092	0.138	0.045	0.139	0.196
2000	U	0.744	0.846	0.890	0.749	0.866	0.911	0.860	0.913	0.941	0.859	0.928	0.951
	E	0.275	0.139	0.190	0.184	0.208	0.259	0.096	0.206	0.264	0.077	0.313	0.363
Saturated model													
500	U	0.371	0.474	0.565	0.344	0.485	0.579	0.453	0.548	0.641	0.409	0.557	0.652
	E	0.074	0.040	0.081	0.054	0.061	0.103	0.055	0.058	0.104	0.045	0.089	0.139
2000	U	0.633	0.778	0.831	0.611	0.776	0.826	0.764	0.862	0.902	0.714	0.856	0.899
	E	0.241	0.118	0.174	0.147	0.173	0.231	0.110	0.151	0.216	0.080	0.241	0.304

Table 6 Power study: model misspecification

		$J=20$						$J=40$					
		α with $\rho=0.3$			α with $\rho=0.6$			α with $\rho=0.3$			α with $\rho=0.6$		
K	I	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$
True = DINA model, Fitted = A-CDM													
3	500	0.113	0.146	0.135	0.081	0.129	0.090	0.075	0.308	0.192	0.071	0.250	0.173
	2000	0.100	0.391	0.310	0.115	0.453	0.383	0.089	0.429	0.236	0.075	0.144	0.134
5	500	0.134	0.279	0.269	0.080	0.262	0.234	0.101	0.620	0.376	0.093	0.368	0.308
	2000	0.296	0.632	0.493	0.186	0.512	0.366	0.114	0.778	0.603	0.145	0.930	0.450
True = A-CDM, Fitted = DINA model													
3	500	0.113	0.146	0.135	0.081	0.129	0.090	0.075	0.308	0.192	0.071	0.250	0.173
	2000	0.100	0.391	0.310	0.115	0.453	0.383	0.089	0.429	0.236	0.075	0.144	0.134
5	500	0.134	0.279	0.269	0.080	0.262	0.234	0.101	0.620	0.376	0.093	0.368	0.308
	2000	0.296	0.632	0.493	0.186	0.512	0.366	0.114	0.778	0.603	0.145	0.930	0.450

0.186 or below with few exceptions when the true model was the DINA model, and the fitted model was the A-CDM, 0.097 or below with few exceptions when verse versa). When the true model was the A-CDM, and the fitted model was the DINA model, the power rates were even lower because the DINA model is simper than the A-CDM.

5 Discussion

A Mantel-Haenszel(MH) statistic proposed by Lim and Drasgow (2019) was evaluated for detecting misspecifications of the latent attribute space in parametric cognitive diagnosis models; that is, the Q-matrix might contain too many or too few latent attributes. (Recall that a misspecified latent attribute space may result in inaccurate parameter estimates that will cause incorrect assessments of examinees' ability.) The proposed MH statistic uses as the levels of the stratification variable the different proficiency classes, with examinees' individual attribute vectors—that identify proficiency class membership—estimated from the data. Simulation studies were conducted for investigating the diagnostic sensitivity of the MH statistic in terms of Type-I-Error rate and power under a variety of testing conditions. Across different sample sizes, test lengths, number of attributes defining the true attribute space, and levels of correlation between the attributes, the MH statistic consistently attained a Type-I-Error rate that was typically close to the nominal $0.05 - \alpha$-level when the data were generated using the true Q-matrix based on the correctly specified latent attribute space. When the data were generated using a Q-matrix based on an under-specified latent attribute space, the MH statistic displayed moderate power in detecting the resulting conditional dependence among test items. In summary, the MH statistic might be a promising tool for uncovering possible misspecifications of the latent attribute space in cognitive diagnosis. Further research is needed to investigate the specific factors that affect the power of the MH statistic; especially, when the latent attribute space has been over-specified (i.e., too many attributes have been included).

References

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*, 123–140.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289.

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

Lim, Y. S., & Drasgow, F. (2019). Conditional independence and dimensionality of nonparametric cognitive diagnostic models: A test for model fit. *Journal of*,. Classification.

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

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institute*, *22*, 719–748.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2015). CDM: Cognitive diagnostic modeling. R package version 3.4–21.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumption of item response theory. *Psychometrika*, *49*, 425–436.

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

Stout, W. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52*, 589–618.

Comparison of Three Unidimensional Approaches to Represent a Two-Dimensional Latent Ability Space

Terry Ackerman, Ye Ma and Edward Ip

Abstract All test data represent the interaction of examinee abilities with individual test items. It has been argued that for most tests these interactions result in, either unintentionally or intentionally, multidimensional response data. Despite this realization, many standardized tests report a single score which follows from fitting a unidimensional model to the response data. This process is justified with the understanding that the response data, when analyzed, say for example by a principal component analysis, have a strong, valid, and content identifiable first component and weaker minor inconsequential components. It is believed that the resulting observed score scale represents primarily a valid composite of abilities that are intended to be measured. This study examines three approaches which estimate unidimensional item and ability parameters based on the parameters obtained from a two-dimensional calibration of the response data. The goal of this study is to compare the results of the different approaches to see which best captures the results of the two-dimensional calibration.

1 Introduction

Testing practitioners are often faced with difficult decisions when they examine the response data from their standardized assessments and the data do not appear to be strictly unidimensional. One option would be to report individual subscores for each dimension if the subscores are indeed reliable and reflect a profile of skills the test

T. Ackerman (✉) · Y. Ma
Department of Psychological and Quantitative Foundations, University of Iowa, Iowa City, IA, USA
e-mail: taackerman@uiowa.edu

Y. Ma
e-mail: ye-ma@uiowa.edu

E. Ip
Wake Forest School of Medicine, Wake Forest University, Bowman Gray Center for Medical Education, 475 Vine Street, Winston-Salem, NC 27101, USA
e-mail: eip@wakehealth.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_18

was designed to measure. Another option, which is commonly followed, would be to calibrate the multidimensional data to fit a unidimensional model and assume that the linear composite being estimated captures the purported skill.

This study compares three different approaches, each of which reduces the two-dimensional latent ability space to a unidimensional space with unidimensional item and ability parameter estimates. The purpose is to identify which approach best recovers characteristics of the underlying two-dimensional model.

2 Three Approaches Investigated

2.1 Approach 1: Projective IRT Model

Ip (2010) and Ip and Chen (2012) developed what they called the Projective IRT (PIRT) and it's unidimensional item parameter approximations. The goal of their work was to provide practitioners with a way to eliminate invalid dimensionality from their item and ability estimates. The approach essentially integrated out the nuisance dimensions resulting in a more valid, albeit dependent, unidimensional 2PL IRT model. The 2PL PIRT model unidimensional item parameters for which θ_1 is the valid dimension are given as

$$a_i^* = \lambda_{logit}\left(a_{i1} + \frac{a_{i2}\rho\sigma_2}{\sigma_1}\right);\ d_i^* = \lambda_{logit} d_i \tag{1}$$

where a* and d* are the PIRT discrimination and difficulty parameters respectively, with

$$\lambda_{logit} = \left[k^2 a_2^2\left(1-\rho^2\right)\sigma_2^2 + 1\right]^{-1/2},\ k = \frac{16\sqrt{3}}{(15\pi)} = 0.588 \tag{2}$$

and where ρ is the correlation between the two dimensions, d_i is the intercept parameter, and a_{i1}, a_{i2}, and σ_1, σ_2 are the discrimination parameter and standard deviations of the θ_1 and θ_2 distributions, respectively.

Ip (2010) demonstrated that response data generated from the 2D MIRT and the (locally dependent) PIRT are empirically indistinguishable. In Ip and Chen (2012) established that the RMSEs between a two-dimensional IRT model and the projected model are quite comparable. They also found that when the MIRT model was miss-specified (i.e., the correct model was say 3D or 5D) that the PIRT model based upon a 2D MIRT solution still yielded a robust performance.

2.2 Approach 2: The Reference Composite

Wang (1986) and Camilli (1992) analytically derived the relationship between unidimensional 2PL item parameter estimates from a 1-D calibration and the underlying 2-D latent space for item *j.* That is, they postulated that the relationship between an underlying two-dimensional item and ability parameters and the estimated item and ability parameters that would result if one were to fit a 2PL IRT model to two-dimensional data. Their unidimensional 2PL model approximations for the discrimination and difficulty parameters are respectively calculated using the formulas

$$\widehat{a}_j = \frac{\boldsymbol{a}'_j \boldsymbol{W}_1}{\sqrt{2.89+\boldsymbol{a}'_j \mathbf{W}_2 \boldsymbol{W}'_2 \boldsymbol{a}_j}};\ \widehat{b}_j = \frac{d_j - a'_j \boldsymbol{\mu}}{\boldsymbol{a}_j \boldsymbol{W}_1} \tag{3}$$

where $\boldsymbol{a}_j$ is the discrimination vector for the M2PL model, d_j is the M2PL difficulty parameter, $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are the first and second standardized eigenvalues of the matrix $L'A'AL$, where $\boldsymbol{A}$ represents the matrix of discrimination parameters for all the items on the test and $L'L = \Omega$. In their formulation the resulting unidimensional ability estimate was estimated as

$$\theta_{uni} = \boldsymbol{\theta}'_i \mathbf{W}_1 \tag{4}$$

where $\boldsymbol{\theta}'_i$ represents the transpose of the two-dimensional ability vector.

Wang (1986) termed the unidimensional composite that would result from fitting a unidimensional model to two-dimensional data as the *reference composite*. The angular direction between the reference composite and the positive θ_1-axis is equal to arccosine of the first element of $\mathbf{W_1}$ and therefore is a function of the test's discrimination parameters and the underlying two-dimensional ability distribution of the examinees. Thus, the direction of the composite scale through the two-dimensional latent ability space is not chosen by the testing practitioner, but rather is an artifact of how well the items are discriminating their various composite skills and how the examinees are distributed throughout the latent ability plane.

2.3 Approach 3: User Specified Composite (USC)

Zhang and Wang (1988) presented a paper in which they provided formulas for a prespecified composite direction defined by θ_α where $\alpha = (\alpha_1, \alpha_2, \ldots \alpha_d)^t$ is a constant vector with non-negative α_js such that $\alpha^t \Sigma \alpha = 1$ (i.e., the $\mathrm{Var}(\theta_\alpha) = 1$). The unidimensional ability θ_α is given as

$$\theta_\alpha = \boldsymbol{\alpha}^t \boldsymbol{\theta} = \sum_{j=1}^{d} \alpha_j \theta_j. \tag{5}$$

Then the unidimensional discrimination parameter, a_i* and difficulty parameter b_i* are given as

$$a_i^* = (1+\sigma_i^{*2})^{-\frac{1}{2}} a_i^t \sum \alpha;\ b_i^* = (1+\sigma_i^{*2})^{-\frac{1}{2}} b_i \tag{6}$$

where $\sigma_i^{*2} = a_i^t \Sigma a_i - \left(a_i^t \Sigma \alpha\right)^2$.

The goal of this research was to allow the user to specify a composite direction an investigate the dimensionality of the test and psychometric properties of observed and true scores.

3 Comparison of Approaches

3.1 Description of Data

To compare the three different mapping approaches several analyses were conducted with both simulated and real data. In this study we used response data from 5000 examinees to a large-scale standardized math assessment test that contained 60 items. The data were fit to a two-dimensional IRT compensatory model Eq. (1) using the R software package mirt (Chalmers, 2012). The estimated item and ability parameters were then used in the formulas above for each of the three approaches to calculate unidimensional parameter estimates.

3.2 Graphical Representation

Using the graphical vector representation developed by Reckase (2009), the sixty item vectors are displayed in Fig. 1 along with a red vector which denotes the reference composite direction which indicates the linear composite that would be estimated if the test data were fit to a unidimensional 2PL IRT model. One can see the vectors for the most part lie in a narrow sector with the reference composite having almost an equal weighting of the two dimensions.

The corresponding TCC and its contour are shown in Fig. 2. The reference composite for this test forms an angle of 45.58° with the positive θ_1-axis.

3.3 True Score Distributions

True score distributions using the estimated 2D-MIRT model parameters and the various unidimensional composite directions were created and are shown in Fig. 3. It appears that the Reference Composite appears to capture most closely the shape

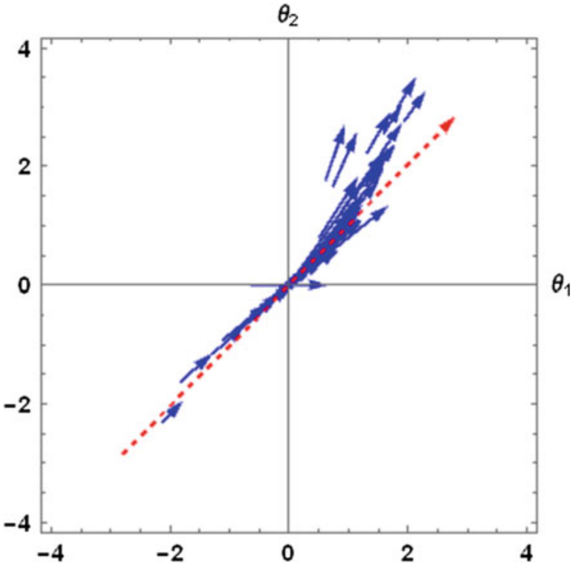


Fig. 1 The vectors for the 60 mathematics items

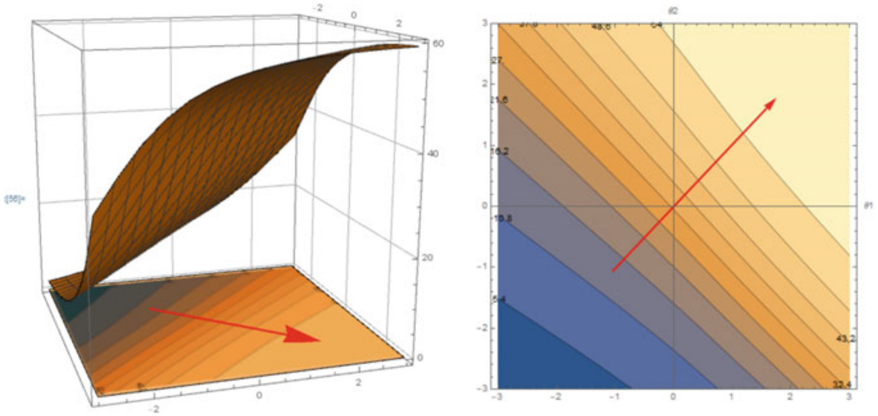


Fig. 2 The TCC surface and its corresponding contour plot with the reference composite direction indicated by the red vector

and distributional characteristics of the true score distribution from the underlying 2D MIRT model.

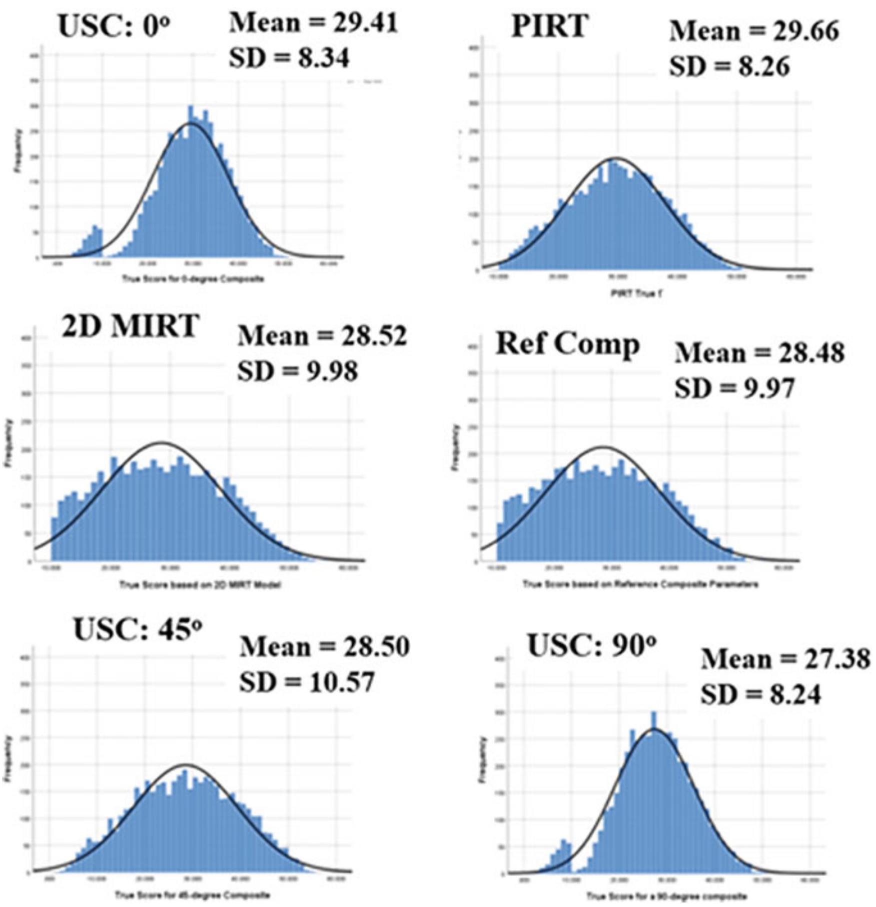


Fig. 3 True score distributions for the estimated 2D MIRT solution and the three different approaches, PIRT, Reference Composite and the USA at 0°, 45° and 90°

3.4 Test Characteristic and Information Curves

The test characteristic curves and the corresponding test information functions for the PIRT, Reference Composite, and the USC unidimensional IRT angular composites at 0°, 30°, 45°, 60° and 90° are displayed in Fig. 4. As might be expected the Reference Composite approach and the USC 45° composite provide the steepest TCC's and the steepest information curves.

The PIRT approach and the USC composites of 0° and 90° provide slightly more than a third maximum information as the Reference Composite and the USC 45° estimates.

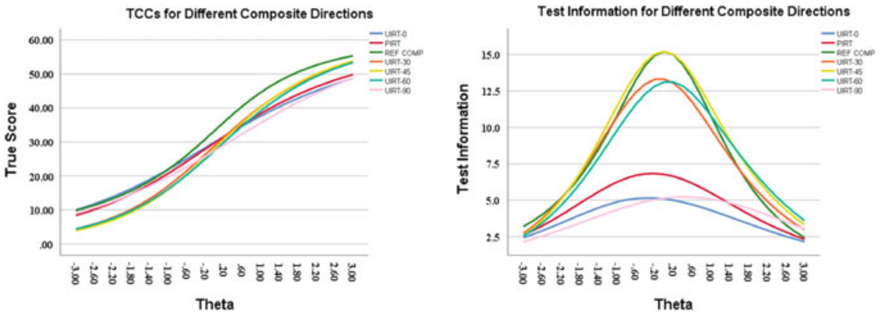


Fig. 4 The TCC's and test information functions using the PIRT, Reference Composite and USC composite directions of 0°, 30°, 45°, 60° and 90°

3.5 RMSE Analysis of True Scores

Another analysis that was completed was to divide the theta scale into twelve equal segments and compute the root mean squared error of the difference between the true score calculated using each unidimensional approach and the true score based upon the two-dimensional MIRT parameters. A plot showing the RMSE for each approach is shown in Fig. 5. The Reference Composite approach clearly provided the smallest RMSE with the PIRT approach providing the largest.

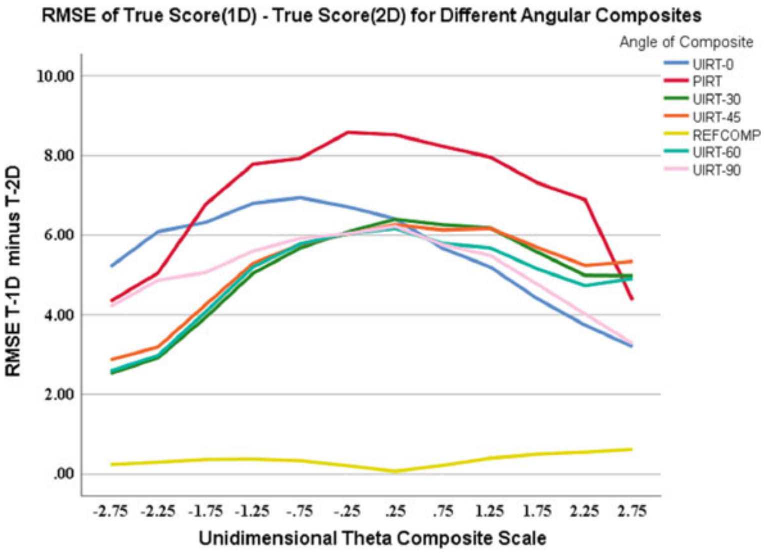
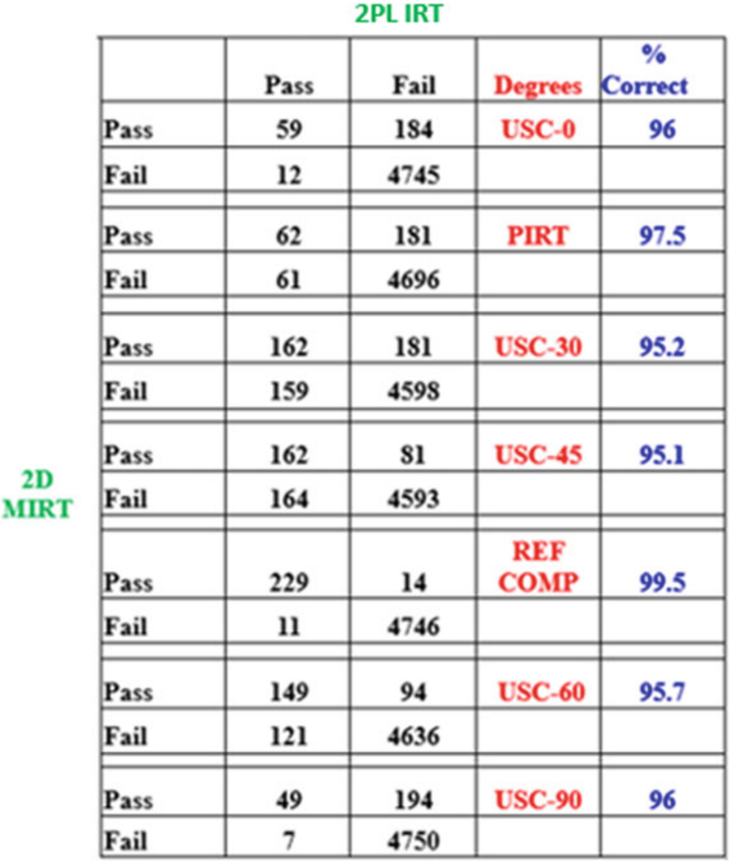


Fig. 5 RMSE of True Score based on unidimensional approaches minus the True Score calculated from the underlying 2D parameters

Table 1 Percent of agreement for passing score decisions between the MIRT 2D parameters and the estimated parameters from the three unidimensional approaches



		2PL IRT			
		Pass	Fail	Degrees	% Correct
2D MIRT	Pass	59	184	USC-0	96
	Fail	12	4745		
	Pass	62	181	PIRT	97.5
	Fail	61	4696		
	Pass	162	181	USC-30	95.2
	Fail	159	4598		
	Pass	162	81	USC-45	95.1
	Fail	164	4593		
	Pass	229	14	REF COMP	99.5
	Fail	11	4746		
	Pass	149	94	USC-60	95.7
	Fail	121	4636		
	Pass	49	194	USC-90	96
	Fail	7	4750		

3.6 Correct Classification Analysis

Another comparison made was to calculate the percent of correct classification decisions That would be made using the true score from the two-dimensional parameters versus the true scores computed from the three approaches for a cut-score of 45. The results appear below in Table 1. Surprisingly the percent agreement is very high for all methods; however, the highest agreement was obtained by the Reference Composite method. Notice that the USC approach for 45° did about as well as the USC approach for 30° and 60° composites.

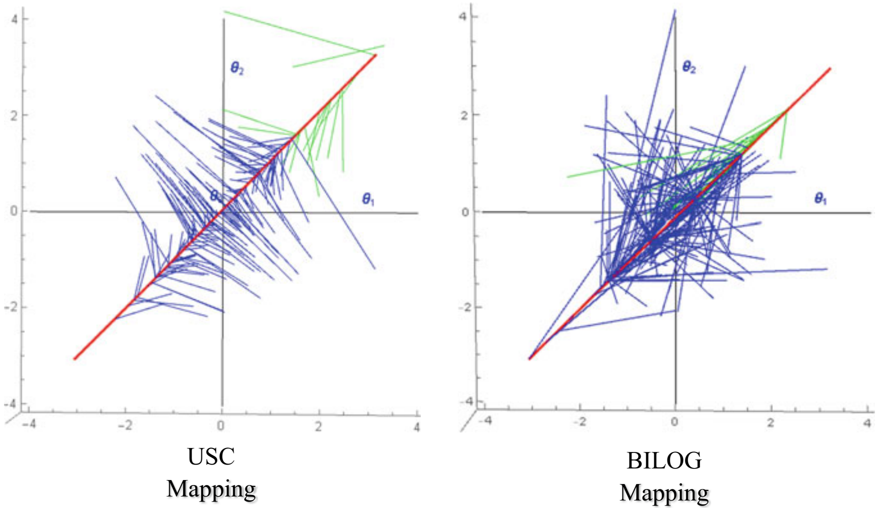


Fig. 6 Mapping of the MIRT (θ_1,θ_2) onto the true score scale using the USC approach and BILOG calibration results

3.7 Comparison of Two-Dimensional to Unidimensional Mappings

A final graphical approach that was conducted to compare the Reference Composite theoretical approach to that obtained from a unidimensional BILOG calibration. Two plots were created in which segments were drawn from examinees estimated (θ_1, θ_2) to their estimated true score calculated using the theoretical Reference Composite formulation and one in which segments were drawn from the estimated (θ_1, θ_2) to a true score based upon a BILOG unidimensional theta and item parameter estimation. The true score scale occupies the reference composite direction. The mappings are shown in Fig. 6. This graph was created for only 100 randomly selected examinees. Green segments indicate examinees which surpassed the cut score of 45. One can clearly see how BILOG was not nearly as consistent in its mapping of the 2-D latent abilities compared with the theoretical approach based upon the 2D parameter estimates.

4 Discussion

The Reference Composite formulation, which was intended to capture a unidimensional calibration of MIRT data, does extremely well for capturing the distribution of true scores produced from the 2D MIRT model. This approach best captured the distribution of true scores based upon the 2D calibration, had the steepest TCC, provided the most test information across the unidimensional theta scale and had the highest percentage of agreement with the 2D data for the cut score that was established.

It should be noted that when the Wang and Zhang composite approach matched the direction of the Reference Composite the approximated 2PL a-parameters were identical, however the b-parameters were noticeably different. Because the PIRT projects onto the first dimensions and the data sets that were examined had reference composite angles at 45.58°, in fairness to the PIRT model, the analyses that were conducted obviously did not favor this model.

References

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement*, *16*(2), 129–147.

Chalmers, P. R. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48,* 1–29.

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology, 63,* 395–415. https://doi.org/10.1177/0146621612462759.

Ip, E. H., & Chen, S. H. (2012). Projective item response model for test-independent measurement. *Applied Psychological Measurement, 36,* 581–601. https://doi.org/10.1177/0146621612452778.

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

Wang, M. (1986, April). *Fitting a unidimensional model to multidimensional response data*. Paper presented at the ONR Contractors Conference, Gatlinburg, TN. https://doi.org/10.1177/014662169101500103.

Zhang, J., & Wang, M. (1998, April). *Relating reported scores to latent traits in a multidimensional test*. Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Comparison of Hyperpriors for Modeling the Intertrait Correlation in a Multidimensional IRT Model

Meng-I Chang and Yanyan Sheng

Abstract Markov chain Monte Carlo (MCMC) algorithms have made the estimation of multidimensional item response theory (MIRT) models possible under a fully Bayesian framework. An important goal in fitting a MIRT model is to accurately estimate the interrelationship among multiple latent traits. In Bayesian hierarchical modeling, this is realized through modeling the covariance matrix, which is typically done via the use of an inverse Wishart prior distribution due to its conjugacy property. Studies in the Bayesian literature have pointed out limitations of such specifications. The purpose of this study is to compare the inverse Wishart prior with other alternatives such as the scaled inverse Wishart, the hierarchical half-t, and the LKJ priors on parameter estimation and model adequacy of one form of the MIRT model through Monte Carlo simulations. Results suggest that the inverse Wishart prior performs worse than the other priors on parameter recovery and model-data adequacy across most of the simulation conditions when variance for person parameters is small. Findings from this study provide a set of guidelines on using these priors in estimating the Bayesian MIRT models.

Keywords Multidimensional item response theory · Fully bayesian model · Markov chain Monte Carlo

1 Introduction

Item response theory (IRT; Lord, 1980) is a modern test theory that has been widely used in educational and psychological measurements (e.g., achievement tests, rating scales, and inventories) as well as other areas such as medical, health sciences,

M.-I. Chang (✉)
Department of Psychology, Philander Smith College, Little Rock, AR 72202, USA
e-mail: mchang@philander.edu

Y. Sheng
Department of Counseling, Quantitative Methods, and Special Education,
Southern Illinois University Carbondale, Carbondale, IL 62901, USA
e-mail: ysheng@siu.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_19

quality-of-life, and marketing research. IRT posits that the probability of a correct response to an item is a mathematical function of person and item parameters (Hemker, Sijtsma, & Molenaar, 1995). This paper focuses on a special case of the multidimensional IRT model, namely, the multi-unidimensional model (Sheng & Wikle, 2007) where the test measures multiple latent traits with each item measuring one of them.

Accurate estimation of model parameters from response data is a central problem in the IRT models. In fact, successful applications of IRT highly rely on finding appropriate procedures for estimating the model parameters (Hambleton, Swaminathan, & Rogers, 1991). Modern computational technology and the development of Markov chain Monte Carlo (MCMC; e.g., Hastings, 1970) methods have made it possible for the IRT estimation methods to gradually shift to a fully Bayesian approach (e.g., Béguin & Glas, 2001; Bolt & Lall, 2003; Chib & Greenberg, 1995; Fox & Glas, 2001; Patz & Junker, 1999). Unlike the conventional marginal maximum likelihood (MML; Bock & Aitkin, 1981) method, the fully Bayesian estimation can avoid unreasonable parameter estimates (e.g., Kim, 2007; Mislevy, 1986; Swaminathan & Gifford, 1983), simultaneously estimate person and item parameters by deriving their joint distribution and allow the uncertainty of estimating items to be incorporated by the uncertainty of estimating person parameters. Fully Bayesian estimation relies on MCMC methods that focus on simulating samples from the posterior distribution, and have been proved useful in practically all aspects of Bayesian inference such as parameter estimation and model comparisons. This study uses the no-U-turn sampler (NUTS; Hoffman & Gelman, 2014), an improvement of an efficient MCMC algorithm named Hamiltonian Monte Carlo (HMC; Duane, Kennedy, Pendleton, & Roweth, 1987). The algorithm gets its name "no-U-turn" sampler because it prevents inefficiencies that would arise from letting the trajectories make a U-turn. NUTS generalizes the notion of the U-turn to high dimensional parameter spaces and estimates when to stop the trajectories before they make a U-turn back toward the starting point. Compared to other MCMC algorithms such as Gibbs sampling (Geman & Geman, 1984), the efficient NUTS requires fewer iterations to converge (e.g., Vehtari, Gelman, & Gabry, 2017).

When estimating multidimensional (Reckase, 1997) or multi-unidimensional IRT models under a fully Bayesian framework, a prior specification is required for the covariance matrix to model the interrelationship among multiple latent traits. The usual practice is to adopt an inverse Wishart (IW) prior distribution (e.g., Sheng & Wikle, 2007) due to its conjugacy property (Barnard, McCulloch, & Meng, 2000). However, in the Bayesian (and not specifically IRT) literature, researchers have called attention to limitations of this prior due to bias resulted from low density around zero for the marginal distribution of variances (Gelman, 2006). To overcome this problem, they have proposed other priors such as the scaled inverse Wishart (O'Malley & Zaslavsky, 2008), the hierarchical half-t (Huang & Wand, 2013), and the separation strategy (SS, Barnard et al., 2000) via the use of the LKJ prior (Lewandowski, Kurowicka, & Joe, 2009). The inverse-Wishart prior has been compared with other priors in non-IRT settings for estimating models such as the multivariate linear model (Alvarez, Niemi, & Simpson, 2014), the growth curve model (Liu, Zhang, & Grimm, 2016), and the multilevel autoregressive model (Schuurman, Grasman, & Hamaker, 2016). Their results generally suggested that the inverse-Wishart prior performed relatively worse in recovering model parameters when the true variance was small even with large sample sizes (e.g., Alvarez et al., 2014). This may hold true with estimating multidimensional IRT (MIRT) models. These priors, however, are typically adopted for the variance-covariance matrix (or hyperparameter) for the unobserved latent variables (instead of that for observed variables as with other studies) in MIRT models. In other words, they are higher-level priors and their effects on estimating such models are not limited to the recovery of the correlation parameter. Given this and that they have not been compared in the IRT framework, the purpose of this study is to investigate the impact of these priors on the posterior inference of the covariance matrix when modeling the interrelationship among multiple latent traits in a multidimensional IRT model. The specific focus is on actual situations where the latent traits have small variances given findings from previous studies with other models.

2 Common Prior Densities for a Covariance Matrix

This section describes common prior specifications that have been considered by researchers in the fully Bayesian estimation for the covariance matrix, $\boldsymbol{\Sigma}$. These priors include the inverse Wishart, scaled inverse Wishart, hierarchical half-t, and LKJ priors.

2.1 Inverse Wishart (IW) Prior

The IW prior is considered as the natural conjugate prior for a covariance matrix of a multivariate normal distribution and can be represented as

$$\boldsymbol{\Sigma} \sim IW(\nu, \Lambda)$$

where ν is a scalar degree of freedom and $\boldsymbol{\Lambda}$ is a positive definite matrix of size $d \times d$. A default approach for the IW prior sets $\boldsymbol{\Lambda} = \mathbf{I}$ and $\nu = d + 1$ where $\mathbf{I}$ is an identity matrix. The IW prior is generally adopted because of its conjugate properties with the normal sampling model and this conjugacy allows it to be easily incorporated into MCMC methods. It, however, suffers from the following problems: (1) The uncertainty of all variance parameters is set by a single degree of freedom parameter, which loses its flexibility to incorporate different amount of prior knowledge to other variance components (Gelman, 2014). (2) When $\nu > 1$, it implies that the scaled inverse-χ^2 prior distribution has an extremely low density near 0, which causes bias in Bayesian inferences for these variances (Gelman, 2006). (3) There is a dependency between the correlations and variances when using the IW prior. Specifically, large variances are associated with correlations with absolute values close to 1 while small variances are associated with correlations near 0 (Tokuda, Goodrich, Van Mechelen, Gelman, & Tuerlinckx 2011). (4) When the true variance is small, the posterior for the variance is biased toward larger values and the correlation is biased toward 0. This bias still exists even with large sample sizes (Alvarez et al., 2014).

2.2 Scaled Inverse Wishart (SIW) Prior

Another approach to model the covariance matrix is using of the scaled inverse Wishart (SIW) prior, which is based on the IW distribution with additional parameters for flexibility (O'Malley & Zaslavsky, 2008). The SIW prior is defined as

$$\mathbf{\Sigma} \sim SIW(\nu, \mathbf{\Lambda}, b_i, \delta_i),$$

where ν is the degrees of freedom, $\mathbf{\Lambda}$ is a positive definite matrix with b_i and δ_i being location and standard deviation vector for the scaling parameters. A hierarchical representation of the SIW prior is $\mathbf{\Sigma} \equiv \Delta\, \mathbf{Q}\, \Delta$ where Δ is a diagonal matrix with $\Delta_{ii} = \xi_i$, and

$$\mathbf{Q} \sim IW(\nu, \mathbf{\Lambda}), \log(\xi_i) \sim N\big(b_{i,}, \delta_i^2\big),$$

where the matrix $\mathbf{Q}$ represents the unscaled covariance matrix distribution, ξ_i parameters are auxiliary parameters to adjust the scale. The SIW prior implies that $\sigma_i = \xi_i, \sqrt{\mathrm{Q}_{ii}}, \Sigma_{ij} = \xi_i \xi_j \sqrt{\mathrm{Q}_{ij}}$, and the correlations $\rho_{ij} = \mathrm{Q}_{ij}/\sqrt{\mathrm{Q}_{ii}\mathrm{Q}_{jj}}$ have the same distribution they have under the inverse Wishart on $\mathbf{Q}$. Gelman and Hill (2007) recommended to set $\mathbf{\Lambda} = \mathbf{I}$ and $\nu = d + 1$ to ensure uniform priors on the correlations as in the IW prior but still can be flexible to incorporate some prior information about standard deviations.

2.3 Hierarchical Half-t Prior

Huang and Wand (2013) proposed a hierarchical approach for the covariance matrix as shown below

$$\mathbf{\Sigma} \sim IW(\nu + d - 1, 2\nu\mathbf{\Omega}),\ \lambda_i \sim \text{Gamma}(\frac{1}{2}, \frac{1}{\xi_i^2}) \text{ with } E(\lambda_i) = \frac{\xi_i^2}{2},$$

where ν again is the degrees of freedom, ξ_i is the scale in the marginal deviations, $\boldsymbol{\Omega}$ is a diagonal matrix with the ith element λ_i. An important advantage of using the hierarchical half-t prior is that it suggests that the standard deviations are distributed as a t distribution that is truncated at 0 to cover only positive values with ν degrees of freedom and ξ_i scale, i.e., $\sigma_i \sim t_\nu^+(0, \xi_i)$, as recommended by Gelman (2006). If ξ_i is set to be large, we can obtain weakly informative priors on the variance and maintain the conjugacy of the prior. In addition, setting $\nu = 2$ implies marginally uniform distribution for the correlation coefficient.

2.4 Separation Strategy via the Use of LKJ Prior

Barnard et al. (2000) proposed a separation strategy (SS), which ensures prior independence between standard deviations and correlations. They decomposed the covariance matrix as Σ =**DRD** where **D** is a diagonal matrix with ith element σ_i and **R** is a correlation matrix with ρ_{ij} on its ith row and jth column. In fact, the Stan manual (Stan Development Team, 2017) recommends a SS approach via the use of the LKJ prior when modeling the covariance matrix of a multivariate normal distribution. The LKJ distribution provides a prior on the correlation matrix **R**, which has a Cholesky factorization $\mathbf{R} = \boldsymbol{LL^T}$ where $\boldsymbol{L}$ is a lower-triangular matrix. This decomposition is numerically more stable and efficient than direct matrix inversion. The LKJ distribution has the density $f\ (\mathbf{R}|\eta) \propto |\mathbf{R}|^{\eta-1}$, with $\eta > 0$. When $\eta = 1$, the LKJ distribution leads to a uniform distribution on correlation matrices, while the magnitude of correlations between components decreases as $\eta \rightarrow \infty$.

3 Methods

Monte Carlo simulations were carried out to compare the four prior specifications for modeling the covariance hyperparameter for the latent person parameters in a multi-unidimensional model. This section describes the model as well as the methodology of the study.

3.1 Model

The two-parameter logistic (2PL) multi-unidimensional model is defined as

$$P\left(Y_{vij} = 1|\theta_{vi}, a_{vj}, b_{vj}\right) = \frac{\exp[a_{vj}\left(\theta_{vi} - b_{vj}\right)]}{1 + \exp[a_{vj}\left(\theta_{vi} - b_{vj}\right)]}, \tag{1}$$

where Y_{vij} is the correct $\left(Y_{vij}=1\right)$ or incorrect $\left(Y_{vij}=0\right)$ response of the ith individual to the jth item in the vth dimension, θ_{vi} is the latent trait parameter for person i on dimension v, a_{vj} is the discrimination parameter of the jth item in dimension v, and b_{vj} is the difficulty parameters of the jth item in dimension v.

3.2 Simulation Procedures

Dichotomous responses of N persons ($N = 100$ or 500) to K items ($K = 40$ or 100) were simulated from the multi-unidimensional model as defined in (1), where $v = 2$. Person parameters $\boldsymbol{\theta}_i = (\theta_{1i}, \theta_{2i})'$ were generated from a bivariate normal distribution such that

$$\boldsymbol{\theta}_i \sim N_2\left(\boldsymbol{\mu}, \sum\right),$$

where $\boldsymbol{\mu} = (0, 0)'$ and $\sum = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. The values of σ_1^2 and σ_2^2 were set to be equal (i.e., 0.01 or 0.1) (e.g., Alvarez et al., 2014; Bouriga & Féron, 2013). The intertrait correlation ρ_{12} was manipulated to be 0.2, 0.5, or 0.8. Item parameters were generated from uniform distributions such that a_{vj} ~ $U(0, 2)$ and b_{vj} ~ $U(-2, 2)$.

To implement the MCMC procedure, the prior for $\boldsymbol{\theta}_i$ was assumed to follow a multivariate normal distribution

$$\boldsymbol{\theta}_i \sim MVN\left(\boldsymbol{\mu}, \sum\nolimits_H\right),$$

where $\boldsymbol{\mu} = (0, 0)'$, and the covariance matrix $\sum_H$ had an IW prior, a SIW prior, a hierarchical half-t prior, or an LKJ prior. The prior specifications for $\sum_H$ are shown in Table 1. These values were adopted to imply marginal non-informative priors on the correlations (Alvarez et al., 2014). In addition, prior densities for a_{vj} and b_{vj} were set following the Bayesian IRT literature, such that $a_{vj} \sim N_{(0,\infty)}$ (0, 1) (e.g., Chang & Sheng, 2016) and $b_{vj} \sim N(0, 1)$ (e.g., Sheng & Wikle, 2007).

Table 1 Parameter values for $\sum_H$ in the simulation study

Prior	Hyperparameter values for prior sampling
IW	$v = 3$, $\Lambda = \mathrm{I}$
SIW	$v = 3$, $\Lambda = \mathrm{I}$, $b_i = 0$, ξ_i ~ $U(0, 100)$
Half-t	$v = 3$, $\xi_i = 100$
LKJ	$\eta = 1$, σ_i ~ Half-Cauchy(0, 2.5)

NUTS was implemented via the use of Stan (Carpenter et al., 2016) where the warm-up stage was set to 500 iterations followed by four chains with 1500 iterations. Convergence of the Markov chains was evaluated using the Gelman-Rubin R statistic (Gelman & Rubin, 1992) with $\hat{R} < 1.1$, suggesting that the chain has converged to the posterior distribution (Brooks & Gelman, 1998). 10 replications for each simulation condition were conducted to avoid erroneous results in estimation due to sampling error.

3.3 Measures of Estimation Accuracy

The accuracy of item and person parameter estimates was evaluated using *bias* and the root mean square error (*RMSE*). *Bias* is defined as

$$bias_{\pi} = \frac{\sum_{j=1}^{n}\left(\hat{\pi}_j - \pi_j\right)}{n}, \tag{2}$$

where π_j (e.g., a_{vj}, b_{vj}, or θ_{vi}) is the true value of an item or person parameter, $\hat{\pi}_j$ is the estimated value of that parameter in the *j*th replication, and n is the total number of replications. If *bias* is close to zero, it suggests that the value of the estimated parameter is close to the true parameter. Also, positive bias suggests that the true parameter is overestimated and a negative bias suggests an underestimation of the true parameter (Dawber, Roger, & Carbonaro, 2009).

The *RMSE* for each item parameter was calculated using the following formula

$$RMSE_{\pi} = \sqrt{\frac{\sum_{j=1}^{n}(\hat{\pi}_j - \pi_j)^2}{n}}, \tag{3}$$

where π_j, $\hat{\pi}_j$, and n are as defined in Eq. (2).

The *RMSE* measures the average squared discrepancy between a set of estimated and true parameters and can be conceived as the amount of variability around a point estimate. In general, a smaller value of the *RMSE* suggests that the more accurate the parameter estimate is. These measures were averaged across items or persons to provide summary information.

3.4 Model Selection

In this study, two fully Bayesian fit indices were considered, namely, the widely available information criterion (WAIC; Watanabe, 2010) and the leave-one-out cross-validation (LOO; Geisser & Eddy, 1979) given that they perform better than other partially Bayesian fit measures such as the deviance information criterion (DIC;

Spiegelhalter, Best, Carlin, & van der Linde, 2002) in the context of IRT model selections (Luo & Al-Harbi, 2017). Similar to any information criteria, smaller values of WAIC and LOO suggest a better model-data adequacy.

4 Results

For the 2PL multi-unidimensional IRT model, the Gelman-Rubin $\hat{R}$ is less than 1.1 for each model parameter using the IW, SIW, Half-t, and LKJ priors under all simulation conditions, suggesting that convergence is potentially achieved. The average *RMSE*s of intertrait correlation (ρ_{12}) parameters, and the average *RMSE*s averaged across items for recovering the discrimination (a_1, a_2), difficulty (b_1, b_2), and averaged across subjects for recovering person ability (θ_1, θ_2),parameters in the 2PL multi-unidimensional model using these four priors are summarized in Fig. 1. The results of the average *RMSE*s for the two dimensions are similar so only the average *RMSE*s of the first dimension are presented in the figure. Also, the results of the average *bias* are not presented since there are no specific patterns. For all item parameters, a_1 and b_1 are used to denote the discrimination and difficulty parameters for items in the first subtest, which are assumed to measure θ_1, and a_2 and b_2 are used to denote the discrimination and difficulty parameters for items in the second subtest, which are assumed to measure θ_2.

4.1 Parameter Recovery

Results on recovering the intertrait correlation parameter show that the *RMSE*s for estimating ρ_{12} using the IW prior tend to be slightly smaller than those using the other priors when $\rho_{12} = 0.2$ and $\sigma^2 = 0.1$. On the other hand, the *RMSE*s for estimating ρ_{12} using the LKJ prior tend to be slightly smaller than those using the other priors when $\rho_{12} = 0.8$ regardless of *N*, *K* or σ^2. In addition, all four priors consistently underestimate ρ_{12} when $\sigma^2 = 0.01$ regardless of *N* or *K*. As the sample size increases, the *RMSE*s for estimating ρ_{12} consistently decrease using all four priors when $\sigma^2 = 0.1$ and using the IW or SIW prior when $\sigma^2 = 0.01$. As the test length increases, the *RMSE*s for estimating ρ_{12} tend to decrease using the four prior specifications for $\mathbf{\Sigma}$ except for the condition where $\rho_{12} = 0.2$ and $\sigma^2 = 0.01$. This pattern, however, is not observed with *bias*. In addition, as the intertrait correlation increases, the *RMSE*s for estimating ρ_{12} tend to increase using all four priors when $\sigma^2 = 0.01$.

For item parameter recovery, the results show that the *RMSE*s for estimating the discrimination and difficulty parameters tend to be slightly larger using the IW prior than those using the other priors when $\sigma^2 = 0.01$. The *RMSE*s for estimating the item parameters (a_1, a_2) and (b_1, b_2) parameters, however, are similar using the four priors when $\sigma^2 = 0.1$. As the sample size increases, the *RMSE*s for estimating the

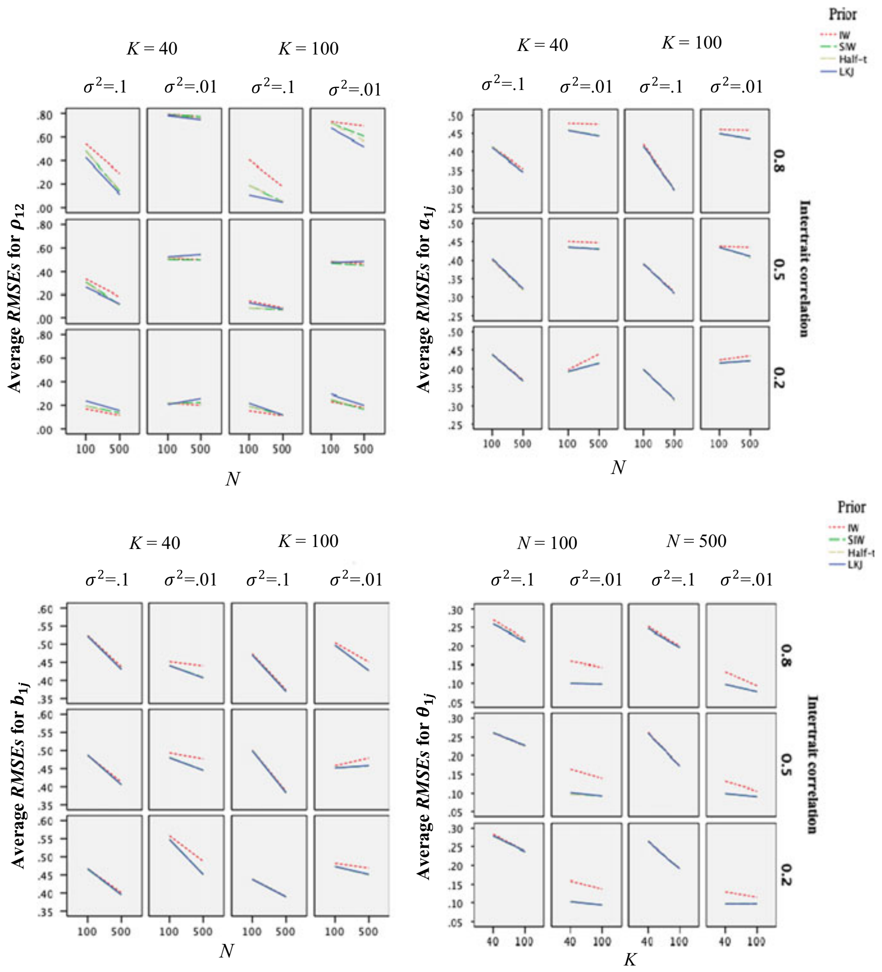


Fig. 1 Average *RMSEs* for recovering the intertrait correlation (ρ_{12}), discrimination (a_1), difficulty (b_1), and person ability (θ_1) parameters under various test conditions in the 2PL multi-unidimensional IRT model

item parameters tend to decrease using all four priors when $\sigma^2 = 0.1$. This pattern, however, is not observed with *bias*. As the test length increases, the *RMSE*s for estimating (a_1, a_2) and (b_1, b_2) do not show a consistent pattern using these four priors. Similarly, there is no clear pattern observed with *bias*. In other words, sample size plays a more crucial role than test length in improving the precision of the item parameter estimates. When comparing the average *RMSE*s for estimating the discrimination and difficulty parameters from different dimensions (i.e., a_1 vs. a_2 and b_1 vs. b_2) under various test conditions, the results indicate that there is no consistent pattern using these four priors.

For person ability parameter recovery, the results show that the *RMSE*s for estimating the person parameter (θ_1, θ_2) tend to be slightly larger using the IW prior than those using the other priors when $\sigma^2 = 0.01$. The *RMSE*s for estimating the person parameters, however, are similar using the four priors when $\sigma^2 = 0.1$. As the test length increases, the *RMSE*s for estimating the (θ_1, θ_2) parameters tend to decrease using all four priors when $\sigma^2 = 0.1$. This pattern, however, is not observed with *bias*. As the sample size increases, the *RMSE*s for estimating (θ_1, θ_2) do not show a consistent pattern using these four priors. Similarly, there is no pattern observed with *bias*. In other words, test length plays a more crucial role than sample size in recovering the person ability parameters. When comparing the average *RMSE*s for estimating the person ability parameters from different dimensions (i.e., θ_1 vs. θ_2) under various test conditions, the results indicate that there is an inconsistent pattern using these four priors.

4.2 Model Selection

Regarding the model-data adequacy, Table 2 summarizes the WAIC and LOO values averaged across 10 replications for the 2PL multi-unidimensional IRT model using the four priors under various simulation conditions. Only the results of $N = 100$ are presented in the table since the results of $N = 500$ are similar. The results show that the WAIC and LOO values are similar using these four priors under various simulation conditions when $\sigma^2 = 0.1$. The WAIC and LOO values, however, tend to be larger using the IW prior than those using the other priors when $\sigma^2 = 0.01$. As for the comparison between the WAIC and LOO, they lead to the same conclusion regarding which model is the best under most of the simulation conditions.

Table 2 WAIC and LOO for the 2PL multi-unidimensional IRT model when $N = 100$ and $\sigma_1^2 = 0.1(0.01)$

N	K	ρ_{12}		IW	SIW	Half-t	LKJ
100	40	0.2	WAIC	4411.06 (4449.92)	4412.78 (4438.38)	4413.37 (4438.36)	4412.77 (4438.28)
			LOO	4413.90 (4451.37)	4414.77 (4438.83)	4415.33 (4438.81)	4414.66 (4438.79)
		0.5	WAIC	4549.85 (4655.33)	4548.11 (4646.90)	4548.26 (4646.60)	4547.85 (4647.08)
			LOO	4552.57 (4656.95)	4549.88 (4647.46)	4550.02 (4647.18)	4549.62 (4647.70)

(continued)

Table 2 (continued)

N	K	ρ_{12}		IW	SIW	Half-t	LKJ
		0.8	WAIC	4484.06 (4501.29)	4484.04 (4490.73)	4483.34 (4490.49)	4483.35 (4490.97)
			LOO	4487.05 (4503.09)	4486.20 (4491.35)	4485.51 (4491.12)	4485.45 (4491.59)
	100	0.2	WAIC	11,286.08 (11,217.34)	11,288.09 (11,201.81)	11,287.71 (11,202.18)	11,288.37 (11,201.82)
			LOO	11,289.62 (11,219.65)	11,291.11 (11,202.67)	11,290.70 (11,203.04)	11,291.34 (11,202.65)
		0.5	WAIC	11,208.28 (11,013.42)	11,208.42 (10,996.07)	11,208.32 (10,995.92)	11,208.66 (10,995.87)
			LOO	11,211.95 (11,015.73)	11,211.48 (10,996.96)	11,211.39 (10,996.82)	11,211.69 (10,996.73)
		0.8	WAIC	11,192.27 (11,280.47)	11,187.64 (11,264.09)	11,187.69 (11,264.71)	11,187.59 (11,264.24)
			LOO	11,195.54 (11,282.80)	11,190.08 (11,265.01)	11,190.15 (11,265.61)	11,189.93 (11,265.14)

5 Conclusions and Discussion

This study compares the IW, SIW, Half-t, and LKJ priors in the performance of parameter recovery for the 2PL multi-unidimensional model through manipulating four factors: intertrait correlation (ρ_{12}), sample size (N), test length (K), and variance (σ^2) for person parameters. When considering the effects of priors, results on the recovery of item and person parameters for the 2PL multi-unidimensional model indicate that the IW prior performs relatively worse than the other priors across most of the simulation conditions when $\sigma^2 = 0.01$. In this particular scenario, from the item parameter recovery point of view, the scaled IW, Half-t, and LKJ priors can be utilized interchangeably to replace the IW prior to have slightly better performance when variance (σ^2) for person parameters is small (i.e., $\sigma^2 = 0.01$). In terms of recovering the intertrait correlation parameter in the 2PL multi-unidimensional model, the IW prior is recommended when multiple latent traits have low correlations (e.g., $\rho_{12} \leq 0.2$), but the LKJ prior is recommended when the traits have high correlations (e.g., $\rho_{12} \geq 0.8$). It is noted that when the true variance for latent person parameters is small ($\sigma^2 = 0.01$), all four priors resulted in poor recovery of the intertrait correlation. This finding is not consistent with Alvarez et al. (2014) who found that all priors work well except for the IW prior when the true variance is small under which the posterior for the variance is biased toward larger values and the correlation is biased toward zero. As pointed out earlier, the four priors are adopted in the multidimensional IRT models as hyperpriors of the covariance hyperparameter for the latent person parameter. This is fundamentally different from previous studies where these priors are used for the covariance parameter for observed variables.

In the IRT setting, small variances for person trait parameters imply that all person trait levels cluster near the mean, which is unusual in the actual (large-scale) testing situations. This study, however, provides empirical evidence that using the IW prior in the 2PL multi-unidimensional model has similar limitations when the person variance is small, which can be useful for situations where the test is given to a more homogenous group of examinees.

In addition, increased sample sizes improve the precision but not the bias in estimating both the discrimination and difficulty parameters, as well as the intertrait correlation when $\sigma^2 = 0.1$. Since increased sample sizes provide more information on estimating items, item parameter estimation improves accordingly. Test length, however, has no consistent effect on the accuracy or bias in estimating item parameters, but tend to improve the precision in estimating the intertrait correlation. In terms of the recovery of the person ability parameters (θ_1 and θ_2), the results suggest that test length has a positive and major effect on estimating θ_1 and θ_2 when $\sigma^2 = 0.1$. More specifically, increased test lengths provide more information on subjects, and therefore, the person ability parameter can be better recovered. These findings are consistent with findings from previous studies on IRT that sample size affects the accuracy of item parameter estimation and test length affects the accuracy of person ability parameter estimation (e.g., Kieftenbeld & Natesan, 2012; Roberts & Thompson, 2011; Sheng, 2010; Swaminathan & Gifford, 1982; Wollack, Bolt, Cohen, & Lee, 2002).

In terms of comparing the performance of these four priors through model-data adequacy, this study shows that the IW prior performs worse than the other prior specifications in model adequacy for data similar to what was considered in this study when $\sigma^2 = 0.01$. This result is consistent with parameter recovery results that the IW prior performs worse than the other prior specifications when $\sigma^2 = 0.01$.

Finally, there are limitations on this study and directions for future studies. It is noted that conclusions are based on simulation conditions considered in the present study and cannot be generalized to other conditions. The present study only considers two sample sizes (i.e., 100 and 500 examinees), two test lengths (i.e., 40 and 100 items), three intertrait correlations (i.e., 0.2, 0.5, and 0.8), two levels of the variance for person parameters (i.e., $\sigma^2 = 0.1$ and 0.01), and equal test items in the two subtests for the 2PL multi-unidimensional model, but for future studies, additional test conditions such as unequal test items or unequal variance for the person parameters can be explored. In addition, the results of this study are based on 10 replications, which are fewer than the minimum number of 25 replications recommended for typical Monte Carlo studies in the IRT-based research (Harwell, Stone, Hsu, & Kirisci, 1996). Due to the fact that MCMC algorithms and model selection procedures are computationally expensive, taking considerable amount of time to execute, it is difficult to go beyond 10 replications in this study for all simulation conditions. Therefore, the results need to be verified with further studies before one can generalize the results to similar conditions. Simulation studies often demonstrate performance under ideal situations. In this case, the true IRT model was known and fit can be assumed to be nearly perfect. Future studies may use these four priors to fit the model to real data and use them for model comparison and selection. In addition, this study only considers two latent dimensions. Future studies can compare the four priors on multi-unidimensional models that have more than two latent dimensions, more general multidimensional, multilevel, or mixture IRT models. Moreover, the findings of the present study are limited to dichotomous models. Models for polytomous responses (e.g., the partial credit or graded response models) should also be explored. Finally, there are a large number of choices for prior distributions or simulated values for the IRT model parameters. Due to that, other prior specifications or simulated values for model parameters a_j, b_j, θ_i, and covariance matrices can also be considered in future studies.

References

Alvarez, I., Niemi, J., & Simpson, M. (2014). Bayesian inference for a covariance matrix. *arXiv preprint* arXiv:1408.4050.

Barnard, J., McCulloch, R., & Meng, X. L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, *10*(4), 1281–1311. Retrieved from http://www.jstor.org/stable/24306780.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66*(4), 541–561. https://doi.org/10.1007/BF02296195.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443–459. https://doi.org/10.1007/BF02293801.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement, 27*(6), 395–414. https://doi.org/10.1177/0146621603258350.

Bouriga, M., & Féron, O. (2013). Estimation of covariance matrices based on hierarchical inverse-Wishart priors. *Journal of Statistical Planning and Inference, 143,* 795–808. https://doi.org/10.1016/j.jspi.2012.09.006.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*(4), 434–455.

Carpenter, B., et al. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1), 1–32. http://dx.doi.org/10.18637/jss.v076.i01.

Chang, M. I., & Sheng, Y. (2016). A comparison of two MCMC algorithms for the 2PL IRT model. In *The Annual Meeting of the Psychometric Society* (pp. 71–79). Springer, Cham.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician, 49*(4), 327–335.

Dawber, T., Rogers, W. T., & Carbonaro, M. (2009). Robustness of Lord's formulas for item difficulty and discrimination conversions between classical and item response theory models. *Alberta Journal of Educational Research, 55*(4), 512–533.

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B, 195,* 216–222. https://doi.org/10.1016/0370-2693(87)91197-X.

Fox, J. P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*(2), 271–288. https://doi.org/10.1007/BF02294839.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association, 74*(365), 153–160.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (Comment on an Article by Browne and Draper). *Bayesian Analysis, 1*(3), 515–533.

Gelman, A. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton: CRC Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge ; New York : Cambridge University Press, 2007.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–472.

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*: Newbury Park, Calif.: Sage Publications.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101–125. https://doi.org/10.1177/014662169602000201.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57*(1), 97–109. https://doi.org/10.1093/biomet/57.1.97.

Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement, 19*(4), 337–352. https://doi.org/10.1177/014662169501900404.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15*(1), 1593–1623.

Huang, A. & Wand, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis, 8*(2), 439–452.

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 36*(5), 399–419. https://doi.org/10.1177/0146621612446170.

Kim, S. H. (2007). Some posterior standard deviations in item response theory. *Educational and Psychological Measurement, 67*(2), 258–279. https://doi.org/10.1177/00131644070670020501.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis, 100*(9), 1989–2001.

Liu, H., Zhang, Z., & Grimm, K. J. (2016). Comparison of inverse Wishart and separation-strategy priors for Bayesian estimation of covariance parameter matrix in growth curve analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(3), 354–367. https://doi.org/10.1080/10705511.2015.1057285.

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

Luo, U., & Al-Harbi, K. (2017). Performances of LOO and WAIC as IRT model selection methods. *Psychological Test and Assessment Modeling, 59*(2), 183–205.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*(2), 177–195. https://doi.org/10.1007/BF02293979.

O'Malley, A., & Zaslavsky, A. (2008). Domain-level covariance analysis for survey data with structured nonresponse. *Journal of the American Statistical Association, 103*(484), 1405–1418. https://doi.org/10.1198/016214508000000724.

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*(2), 146–178. https://doi.org/10.3102/10769986024002146.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25–36. https://doi.org/10.1177/0146621697211002.

Roberts, J., & Thompson, V. (2011). Marginal maximum a posteriori item parameter estimation for the generalized graded unfolding model. *Applied Psychological Measurement, 35*(4), 259–279. https://doi.org/10.1177/0146621610392565.

Schuurman, N. K., Grasman, R. P. P. P., & Hamaker, E. L. (2016). A comparison of inverse-Wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivariate Behavioral Research, 51*(2–3), 185–206. https://doi.org/10.1080/00273171.2015.1065398.

Sheng, Y. (2010). A sensitivity analysis of Gibbs sampling for 3PNO IRT models: Effects of prior specifications on parameter estimates. *Behaviormetrika, 37*(2), 87–110. https://doi.org/10.2333/bhmk.37.87.

Sheng, Y., & Wikle, C. K. (2007). Comparing multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement, 67*(6), 899–919. https://doi.org/10.1177/0013164406296977.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 583–639.

Stan Development Team. (2017). *Stan modeling language users guide and reference manual*, Version 2.15.0.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7*(3), 175–191. https://doi.org/10.2307/1164643.

Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. *New Horizon Testing*, 13–30. https://doi.org/10.1016/b978-0-12-742780-5.50009-3.

Tokuda, T., Goodrich, B., Van Mechelen, I., Gelman, A., & Tuerlinckx, F. (2011). *Visualizing distributions of covariance matrices*. Unpublished manuscript. http://www.stat.columbia.edu/gelman/research/unpublished/Visualization.pdf.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4.

Watanabe, S. (2010, December). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y. S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 26*(3), 339–352. https://doi.org/10.1177/0146621602026003007.

On Extended Guttman Condition in High Dimensional Factor Analysis

Kentaro Hayashi, Ke-Hai Yuan and Ge (Gabriella) Jiang

Abstract It is well-known that factor analysis and principal component analysis often yield similar estimated loading matrices. Guttman (Psychometrika 21:273–285, 1956) identified a condition under which the two matrices are close to each other at the population level. We discuss the matrix version of the Guttman condition for closeness between the two methods. It can be considered as an extension of the original Guttman condition in the sense that the matrix version involves not only the diagonal elements but also the off-diagonal elements of the inverse matrices of variance-covariances and unique variances. We also discuss some implications of the extended Guttman condition, including how to obtain approximate estimates of the inverse of covariance matrix under high dimensions.

Keywords High dimensions · Principal components · Unique variances

1 Factor Analysis and Principal Component Analysis

Factor analysis (FA) and principal component analysis (PCA) are frequently used multivariate statistical methods for data reduction. In FA (Anderson, 2003; Lawley & Maxwell, 1971), the p-dimensional mean-centered vector of the observed variables $\boldsymbol{y}_i$, $i = 1, \ldots, n$, is linearly related to a m-dimensional vector of latent factors $\boldsymbol{f}_i$

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, Corbett Family Hall, Notre Dame,
IN 46556, USA
e-mail: kyuan@nd.edu

G. (Gabriella) Jiang
Department of Educational Psychology, University of Illinois at Urbana-Champaign,
1310 South 6th Street, Campaign, Urbana-Champaign, IL 61820, USA
e-mail: gejiang2@illinois.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_20

via $\boldsymbol{y}_i = \boldsymbol{\Lambda} \boldsymbol{f}_i + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m)$ is a $p \times m$ matrix of factor loadings (with $p > m$), and $\boldsymbol{\varepsilon}_i$ is a p-dimensional vector of errors. Typically for the orthogonal factor model, the three assumptions are imposed: (i) $\boldsymbol{f}_i \sim N_m(\mathbf{0}, \boldsymbol{I}_m)$; (ii) $\boldsymbol{\varepsilon}_i \sim N_p(\mathbf{0}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is a diagonal matrix with positive elements on the diagonal; (iii) $Cov(\boldsymbol{f}_i, \boldsymbol{\varepsilon}_i) = \mathbf{0}$. Then, under these three assumptions, the covariance matrix of $\boldsymbol{y}_i$ is given by $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$. If $\boldsymbol{y}_i$ is standardized, $\boldsymbol{\Sigma}$ is a correlation matrix.

Let $\boldsymbol{\Lambda}^+ = (\boldsymbol{\lambda}_1^+, \ldots, \boldsymbol{\lambda}_m^+)$ be the $p \times m$ matrix whose columns are the standardized eigenvectors corresponding to the first m largest eigenvalues of $\boldsymbol{\Sigma}$; $\boldsymbol{\Omega} = diag(\boldsymbol{\omega})$ be the $m \times m$ diagonal matrix whose diagonal elements $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_m)'$ are the first m largest eigenvalues of $\boldsymbol{\Sigma}$; and $\boldsymbol{\Omega}^{1/2}$ be the $m \times m$ diagonal matrix whose diagonal elements are the square root of those in $\boldsymbol{\Omega}$. Then principal components (PCs) (c.f., Anderson, 2003) with m elements are obtained as $\boldsymbol{f}_i^* = \boldsymbol{\Lambda}^{+\prime}\boldsymbol{y}_i$. Clearly, the PCs are uncorrelated with a covariance matrix $\boldsymbol{\Lambda}^{+\prime}\boldsymbol{\Sigma}\boldsymbol{\Lambda}^+$. When m is properly chosen, there exists $\boldsymbol{\Sigma} \approx \boldsymbol{\Lambda}^+\boldsymbol{\Omega}\boldsymbol{\Lambda}^{+\prime} = \boldsymbol{\Lambda}^*\boldsymbol{\Lambda}^{*\prime}$, where $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}^+\boldsymbol{\Omega}^{1/2}$ is the $p \times m$ matrix of PCA loadings.

2 Closeness Conditions Between Factor Analysis and Principal Component Analysis

It has been well-known that FA and PCA often yield approximately the same results, especially their estimated loading matrices $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\Lambda}}^*$, respectively (e.g., Velicer & Jackson, 1990). Conditions under which the two matrices are close to each other are of substantial interest. At the population level, two such conditions identified by Guttman (1956) and Schneeweiss (1997) are among the most well-known.

2.1 Guttman Condition

Consider the factor analysis model $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is a diagonal unique variance matrix, with $(\boldsymbol{\Sigma}^{-1})_{jj} = \sigma^{jj}$ and $(\boldsymbol{\Psi})_{jj} = \psi_{jj}$, $j = 1, \ldots, p$. Let m be the number of common factors, Guttman (1956; See also Theorem 1 of Krijnen, 2006) has shown that if $m/p \to 0$ as $p \to \infty$, then $\psi_{jj}\sigma^{jj} \to 1$ for almost all j. Here, "for almost all j" means $\lim_{p\to\infty} \#\{j : \psi_{jj}\sigma^{jj} < 1\}/p = 0$. That is, the number of j that satisfies $\psi_{jj}\sigma^{jj} < 1$ is ignorable as p goes to infinity.

2.2 Schneeweiss Condition

The closeness condition between the loading matrix from FA and that from PCA by Schneeweiss and Mathes (1995) and Schneeweiss (1997) is $ev_m(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \to$

∞, where $ev_k(\boldsymbol{A})$ is the k-th largest eigenvalue of a square matrix $\boldsymbol{A}$. Obviously, $ev_m(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})$ is the smallest eigenvalue of $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$.

Related with the Schneeweiss condition, Bentler (1976) parameterized the correlation structure of the factor model as $\boldsymbol{\Psi}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{-1/2} = \boldsymbol{\Psi}^{-1/2}\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1/2} + \boldsymbol{I}_p$ and showed that, under this parameterization, a necessary condition for $ev_m(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) = ev_m(\boldsymbol{\Psi}^{-1/2}\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1/2}) \to \infty$ is that as p increases, the sum of squared loadings on each factor has to go to infinity ($\boldsymbol{\lambda}_k'\boldsymbol{\lambda}_k \to \infty$, $k = 1, \dots, m$, as $p \to \infty$).

2.3 *Relationship Between Guttman and Schneeweiss Conditions*

The relationship between Guttman and Schneeweiss conditions is summarized in Table 1. Schneeweiss condition ($ev_m(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \to \infty$) is sufficient for Guttman condition ($m/p \to 0$ as $p \to \infty$) (Krijnen, 2006, Theorem 3). What we would like is for the converse ($m/p \to 0$ as $p \to \infty \Rightarrow ev_m(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \to \infty$) to hold in practical applications, as to be discussed in the next section.

First, the condition of $m/p \to 0$ as $p \to \infty$ is sufficient for $\psi_{jj}\sigma^{jj} \to 1$ for almost all j (Guttman, 1956; Krijnen, 2006, Theorem 1). Also, $\psi_{jj}\sigma^{jj} \to 1$ for all j implies $ev_m(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \to \infty$ (Krijnen, 2006, Theorem 4). Here, "$\psi_{jj}\sigma^{jj} \to 1$ for all j" is slightly stronger than "$\psi_{jj}\sigma^{jj} \to 1$ for almost all j." However, in practice, it seems reasonable to assume that the number of loadings on every factor increases with p proportionally, as stated in Bentler (1976). Then the condition of $m/p \to 0$ as $p \to \infty$ becomes equivalent to $ev_m(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \to \infty$. That is, Guttman and Schneeweiss conditions become interchangeable.

Table 1 Relationships among conditions and results

Condition(s)	Result	Source
1. $m/p \to 0$ as $p \to \infty$	$\psi_{jj}\sigma^{jj} \to 1$ for almost all j	Guttman (1956), Krijnen (2006, Thm 1)
2. $\psi_{jj}\sigma^{jj} \to 1$ for almost all j; $ev_m(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) > c > 0$	$m/p \to 0$ as $p \to \infty$	Krijnen (2006, Thm 2)
3. $ev_m(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \to \infty$	$m/p \to 0$ as $p \to \infty$	Krijnen (2006, Thm 3), Hayashi and Bentler (2000, Obs 8b)
4. $ev_m(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \to \infty$	$\psi_{jj}\sigma^{jj} \to 1$ for all j	Krijnen (2006, Thm 4), Hayashi and Bentler (2000, after Obs 7)
5. $\psi_{jj}\sigma^{jj} \to 1$ for all j	$ev_m(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \to \infty$	Krijnen (2006, Thm 4)
6. $ev_k(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \to \infty, k = 1, \dots, r$; $ev_k(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) < C < \infty, k = r+1, \dots, m$; $p \to \infty$, m fixed	$\psi_{jj}\sigma^{jj} \to 1$ for almost all j	Krijnen (2006, Thm 5)
7. $ev_m(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \to \infty$	$\boldsymbol{\Psi}^{-1} - \boldsymbol{\Sigma}^{-1} \to \boldsymbol{0}$	

Notes (i) 2 is a partial converse of 1; (ii) 5 is the converse of 4; (iii) 7 is the matrix version of 4

3 Extended Guttman Condition

By far the most important consequence of the Schneeweiss condition is that, when $ev_m(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \to \infty$, the second term in the right-hand side of the Sherman-Morrison-Woodbury formula (see, e.g., Chap. 16 of Harville, 1997):

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}(\boldsymbol{I}_m + \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1} \tag{1}$$

vanishes, so that

$$\boldsymbol{\Psi}^{-1} - \boldsymbol{\Sigma}^{-1} \to \mathbf{0} \quad \text{as} \quad ev_m(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \to \infty \tag{2}$$

As we noted in the previous section, the condition of $m/p \to 0$ as $p \to \infty$ can be equivalent to $ev_m(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \to \infty$ in practical applications. Therefore, we have $\boldsymbol{\Psi}^{-1} - \boldsymbol{\Sigma}^{-1} \to \mathbf{0}$ under high dimensions with a large p. We call $\boldsymbol{\Psi}^{-1} - \boldsymbol{\Sigma}^{-1} \to \mathbf{0}$ the extended Guttman condition. It is an extension of the original Guttman condition in the sense that $\psi_{jj}\sigma^{jj} \to 1$ can be expressed as $\psi_{jj}^{-1} - \sigma^{jj} \to 0$, as long as ψ_{jj} is bounded above ($\psi_{jj} \le \psi_{\text{sup}} < \infty$).

Note that there exists a similar identity for the FA model:

$$\boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1} \tag{3}$$

(see, e.g., Hayashi & Bentler, 2001). Clearly, as $ev_m(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \to \infty$, not only the second term on the left-hand side of Eq. (3) but the second term on the right-hand side of Eq. (3) vanishes.

As we have just seen, the extended Guttman condition is a direct consequence of the Schneeweiss condition. Because $\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}(\boldsymbol{I}_m + \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1} < \boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}$ and $\boldsymbol{I}_m + \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ is only slightly larger than $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ when $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ is large in the sense that $ev_m(\boldsymbol{I}_m + \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) = ev_m(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) + 1$, the speed of convergence in $\boldsymbol{\Psi}^{-1} - \boldsymbol{\Sigma}^{-1} \to \mathbf{0}$ is approximately at the rate of the reciprocal of smallest eigenvalues of $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$, that is, of $1/ev_m(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})$.

4 Approximation of the Inverse of the Covariance Matrix

An important point to note here is that the original Guttman condition of $\psi_{jj}\sigma^{jj} \to 1$ (for almost all j) has to do with only the diagonal elements of $\boldsymbol{\Psi}$ (or $\boldsymbol{\Psi}^{-1}$) and $\boldsymbol{\Sigma}^{-1}$, while $\boldsymbol{\Psi}^{-1} - \boldsymbol{\Sigma}^{-1} \to \mathbf{0}$ involves both the diagonal and the off-diagonal elements of the matrices. It justifies the interchangeability of $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Psi}^{-1}$ as $ev_m(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \to \infty$, or assuming that the number of loadings on every factor increases with p proportionally, as $m/p \to 0$ with $p \to \infty$. The important implication is that all the off-diagonal elements of $\boldsymbol{\Sigma}^{-1}$ approach zero in the limit. Thus, it is a result of sparsity of the off-diagonal elements of the inverted covariance (correlation) matrix in high dimensions.

One of the obvious advantages of being able to approximate $\boldsymbol{\Sigma}^{-1}$ by $\boldsymbol{\Psi}^{-1}$ in high dimensions is that the matrix of unique variances $\boldsymbol{\Psi}$ is a diagonal matrix and thus it can be inverted only with p operations. Note that, in general, the inversion of a p –dimensional square matrix requires operations of order $O(p^3)$ (see, e.g., Pourahmadi, 2013, p. 121).

Consequently, the single most important application of the extended Guttman condition is to approximate the inverse of the covariance matrix $\boldsymbol{\Sigma}^{-1}$ by $\boldsymbol{\Psi}^{-1}$ under high dimensions. This implication is very important because $\boldsymbol{\Sigma}^{-1}$ is involved in the quadratic form for the log likelihood function of the multivariate normal distribution. Even if $\boldsymbol{\Sigma}$ is positive definite so that $\boldsymbol{\Sigma}^{-1}$ exists in the population, the inverse $\boldsymbol{S}^{-1}$ of the sample covariance matrix $\boldsymbol{S}$ does not exist under high dimensions when $p > n$. When $\boldsymbol{S}^{-1}$ does not exist, we cannot estimate $\boldsymbol{\Psi}^{-1}$ under the FA model using the generalized least squares (GLS) or the maximum likelihood (ML) method, without resorting to certain regularization method(s), either. Thus, a natural choice would be to employ the unweighted least square (ULS) estimation method that minimizes the fit function of $F_{ULS}(\boldsymbol{S}, \boldsymbol{\Sigma}) = tr\{(\boldsymbol{S} - \boldsymbol{\Sigma})^2\}$, which does not require to compute $\boldsymbol{S}^{-1}$ or the estimate of $\boldsymbol{\Sigma}^{-1}$. Note that $1 - 1/s^{jj}$, a common initial value for the j-th communarity cannot be used because it requires the computation of $\boldsymbol{S}^{-1}$. Then, we can use the value of 1 as the initial communality estimates. In this case, the initial solution is identical to PCA.

Alternatively, when p is huge, we can employ the following "approximate" FA model with equal unique variances (e.g., Hayashi & Bentler, 2000), using standardized variables, that is, applying to the correlation matrix:

$$\boldsymbol{\Sigma} \approx \boldsymbol{\Lambda}^{*}\boldsymbol{\Lambda}^{*'} + k\boldsymbol{I}_p, \tag{4}$$

with a positive constant k. Note that this model is also called the probabilistic PCA in statistics (Tipping & Bishop, 1999). Use of the FA model with equal unique variances seems reasonable, because the eigenvectors of $(\boldsymbol{\Sigma} - k\boldsymbol{I}_p)$ are the same as the eigenvectors of $\boldsymbol{\Sigma}$, and the eigenvalues of $(\boldsymbol{\Sigma} - k\boldsymbol{I}_p)$ are smaller than the eigenvalues of $\boldsymbol{\Sigma}$ by only the constant of k. Thus, the FA model with equal unique variances is considered as a variant of the PCA, and, the loading matrices between the FA and the PCA approach the same limit values as $ev_m(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \to \infty$, or they become essentially equivalent, under high dimensions.

In Eq. (4), let $\boldsymbol{\Psi}^{*} = k\boldsymbol{I}_p$, then $\boldsymbol{\Psi}^{*-1} = k^{-1}\boldsymbol{I}_p$. Thus, we can use $\boldsymbol{\Psi}^{*-1}$ as quick and fast approximation for $\boldsymbol{\Psi}^{-1}$. The natural estimator of k is the MLE for k given $\boldsymbol{\Lambda}^{*}$ (Tipping & Biship, 1999):

$$\hat{k} = \frac{1}{p - m} \sum_{j=m+1}^{p} ev_j(\boldsymbol{S}). \tag{5}$$

However, a more practical method seems as follows: Once estimating $\boldsymbol{\Psi}^* = k\boldsymbol{I}_p$ by $\hat{\boldsymbol{\Psi}}^* = \hat{k}\boldsymbol{I}_p$, we can compute loadings $\hat{\Lambda}^*$ using the eigenvalues and eigenvectors of $(\boldsymbol{S}-\hat{k}\boldsymbol{I}_p)$ and find the estimates of $\boldsymbol{\Psi}^*$ as $\hat{\boldsymbol{\Psi}}^* = diag(\boldsymbol{S}-\hat{\Lambda}^*\hat{\Lambda}^{*\prime})$. Note that $\hat{\boldsymbol{\Psi}}^*$ is no longer a constant times the identity matrix. Now, invoke the estimator version of the extended Guttman condition $\hat{\boldsymbol{\Psi}}^{*-1} - \hat{\boldsymbol{\Sigma}}^{-1} \approx \boldsymbol{0}$ to find the approximate estimator $\hat{\boldsymbol{\Sigma}}^{-1}$ of $\boldsymbol{\Sigma}^{-1}$.

5 Illustration

The compound symmetry correlation structure is expressed as $\boldsymbol{\Sigma} = (1-\rho)\boldsymbol{I}_p + \rho\mathbf{1}_p\mathbf{1}_p'$ with a common correlation ρ, $0 < \rho < 1$. Obviously, it is a one-factor model with the vector of factor loadings $\boldsymbol{\lambda}_1 = \sqrt{\rho}\mathbf{1}_p$ and the diagonal matrix unique variances $\boldsymbol{\Psi} = (1-\rho)\boldsymbol{I}_p$. Because the first eigenvalue and the corresponding standardized eigenvector of $\boldsymbol{\Sigma} = (1-\rho)\boldsymbol{I}_p + \rho\mathbf{1}_p\mathbf{1}_p'$ are $\omega_1 = 1 + (p-1)\rho$ and $\boldsymbol{\lambda}_1^+ = (1/\sqrt{p})\mathbf{1}_p$, respectively, the first PC loading vector is

$$\boldsymbol{\lambda}_1^* = \boldsymbol{\lambda}_1^+\sqrt{\omega_1} = (1/\sqrt{p})\sqrt{1+(p-1)\rho}\cdot\mathbf{1}_p = \sqrt{1/p+(1-1/p)\rho}\cdot\mathbf{1}_p, \qquad (6)$$

which approaches the vector of factor loadings $\boldsymbol{\lambda}_1 = \sqrt{\rho}\mathbf{1}_p$ with $m/p = 1/p \to 0$ and $p \to \infty$. The remaining $p-1$ eigenvalues are $\omega_2 = \ldots = \omega_p = 1-\rho$. Thus, obviously, the constant k in the FA model with equal unique variances is $k = 1-\rho$. Note that the Schneeweiss condition also holds

$$\boldsymbol{\lambda}_1'\boldsymbol{\Psi}^{-1}\boldsymbol{\lambda}_1 = (\sqrt{\rho}\mathbf{1}_p)'\{(1/(1-\rho))\boldsymbol{I}_p\}(\sqrt{\rho}\mathbf{1}_p) = p\cdot\rho/(1-\rho) \to \infty \qquad (7)$$

with $m/p = 1/p \to 0$ as $p \to \infty$. The inverse of the correlation matrix is:

$$\begin{aligned}\boldsymbol{\Sigma}^{-1} &= \boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{\lambda}_1(1+\boldsymbol{\lambda}_1'\boldsymbol{\Psi}^{-1}\boldsymbol{\lambda}_1)^{-1}\boldsymbol{\lambda}_1'\boldsymbol{\Psi}^{-1}\\ &= (\frac{1}{1-\rho})\boldsymbol{I}_p - (\frac{1}{1-\rho})\boldsymbol{I}_p\cdot(\sqrt{\rho}\mathbf{1}_p)\cdot(1+\frac{\rho}{1-\rho}\cdot p)^{-1}\cdot(\sqrt{\rho}\mathbf{1}_p')\cdot(\frac{1}{1-\rho})\boldsymbol{I}_p\\ &= (\frac{1}{1-\rho})\boldsymbol{I}_p - (\frac{\rho}{1-\rho})(\frac{1}{(1-\rho)+\rho\cdot p})(\mathbf{1}_p\mathbf{1}_p') \to (\frac{1}{1-\rho})\boldsymbol{I}_p = \boldsymbol{\Psi}^{-1} \qquad (8)\end{aligned}$$

with $m/p = 1/p \to 0$ as $p \to \infty$.

For example, it is quite easy to show that if $\rho = 0.5$, then for $p = 10$, the diagonal elements of the inverse of the compound symmetry correlation structure are 2 — 1/5.5 = 1.818 and the off-diagonal elements are −1/5.5 = –0.182. At $p = 100$, the diagonal and the off-diagonal elements become 2 — 1/50.5 = 1.980 and −1/50.5 = –0.0198, respectively. Furthermore, at $p = 1000$, the diagonal and the off-diagonal elements become 2 — 1/500.5 = 1.998 and −1/500.5 = –0.001998. Again, we see the off-diagonal elements of $\boldsymbol{\Sigma}^{-1}$ approaching 0 as p increases. Also, the diagonal elements of $\mathbf{\Sigma}^{-1}$ approach 2, which are the value of the inverse of the unique variances in the FA model.

6 Discussion

We discussed the matrix version of the Guttman condition for closeness between FA and PCA. It can be considered as an extended Guttman condition in the sense that the matrix version involves not only the diagonal elements but also the off-diagonal elements of the matrices $\mathbf{\Sigma}^{-1}$ and $\mathbf{\Psi}^{-1}$. Because $\mathbf{\Psi}^{-1}$ is a diagonal matrix, the extended Guttman condition implies that the off-diagonal elements of $\mathbf{\Sigma}^{-1}$ approach zero as the dimension increases. We showed how the phenomenon happens with the compound symmetry example in the Illustration section. We also discussed some implications of the extended Guttman condition, which include the ease of inverting $\mathbf{\Psi}$ compared against inverting $\mathbf{\Sigma}$. Because the ULS estimation method does not involve any inversion of either the sample covariance matrix $\boldsymbol{S}$ or the estimated model implied population covariance matrix $\hat{\mathbf{\Sigma}}$, the ULS should be the estimation of choice when sample size n is smaller than the number of variables p. Furthermore, we proposed a simple method to approximate $\mathbf{\Sigma}^{-1}$ by $\mathbf{\Psi}^{-1}$ using the FA model with equal unique variances, or equivalently, the probabilistic PCA model.

Some other implications of the extended Guttman condition (especially with respect to algorithms) are as follows: First of all, suppose we add the $(p+1)$th variable at the end of already existing p variables. Then, while the values of σ^{jj}, $j = 1, \ldots, p$, can change, ψ_{jj}^{-1}, $j = 1, \ldots, p$, remain unchanged. Thus, with the extended Guttman condition, only one additional element needs to be computed.

Another implication is on the ridge estimator, which is among the methods to deal with singularity of $\boldsymbol{S}$ or the estimator of its covariance matrix by introducing some small bias term (see e.g., Yuan & Chan, 2008, 2016). Warton (2008, Theorem 1) showed that the ridge estimator of the covariance (correlation) matrix $\hat{\mathbf{\Sigma}}_\eta{=}\eta\hat{\mathbf{\Sigma}}+(1-\eta)\boldsymbol{I}_p$ (with the tuning parameter η) is the maximum penalized likelihood estimator with the penalty term proportional to $-tr(\mathbf{\Sigma}^{-1})$. Unfortunately, as the dimension p increases (or the ratio p/n increases), it becomes more difficult to obtain the inverse of the covariance matrix. Therefore, in high dimensions, it is not practical to express the ridge estimator of the covariance matrix in the form of the maximum penalized likelihood with the penalty term involving $-tr(\mathbf{\Sigma}^{-1})$. This naturally leads to employing an "approximate" maximum penalized likelihood with the penalty term approximately proportional to $-tr(\mathbf{\Psi}^{-1})$ in place of the penalty term proportional to $-tr(\mathbf{\Sigma}^{-1})$, assuming the factor analysis model, when the dimension p is large.

We are aware that, perhaps except approximations of the inverse of covariance matrix, the majority of implications that we discussed in this article may be of limited practical utility. For example, because the original Guttman condition, the Schneeweiss condition, and the extended Guttman condition are all conditions for closeness between FA and PCA, we can simply employ PCA as an approximation to FA when the conditions hold. Also, we did not discuss regularized FA with L1 regularization here, which in itself is a very interesting topic. Yet, we think the implications we discussed are still of theoretical interest that should continue to be studied. The compound symmetry example used in the Illustration is probably only an approximation to the real world. We will need to do an extensive simulation to come up with some empirical guidelines regarding how to best apply the theoretical results in practice.

Acknowledgements The authors are thankful to Dr. Dylan Molenaar's comments. Ke-Hai Yuan's work was supported by the National Science Foundation under Grant No. SES-1461355.

References

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). New York: Wiley.

Bentler, P. M. (1976). Multistructure statistical model applied to factor analysis. *Multivariate Behavioral Research, 11,* 3–15.

Guttman, L. (1956). Best possible systematic estimates of communalities. *Psychometrika, 21,* 273–285.

Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*. New York: Springer.

Hayashi, K., & Bentler, P. M. (2000). On the relations among regular, equal unique variances and image factor analysis. *Psychometrika, 65,* 59–72.

Hayashi, K., & Bentler, P. M. (2001). The asymptotic covariance matrix of maximum-likelihood estimates in factor analysis: The case of nearly singular matrix of estimates of unique variances. *Linear Algebra and its Applications, 321,* 153–173.

Krijnen, W. P. (2006). Convergence of estimates of unique variances in factor analysis, based on the inverse sample covariance matrix. *Psychometrika, 71,* 193–199.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method* (2nd ed.). New York: American Elsevier.

Pourahmadi, M. (2013). *High-dimensional covariance estimation*. New York: Wiley.

Schneeweiss, H. (1997). Factors and principal components in the near spherical case. *Multivariate Behavioral Research, 32,* 375–401.

Schneeweiss, H., & Mathes, H. (1995). Factor analysis and principal components. *Journal of Multivariate Analysis, 55,* 105–124.

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B, 61,* 611–622.

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.

Warton, D. I. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *Journal of the American Statistical Association, 103,* 340–349.

Yuan, K.-H., & Chan, W. (2008). Structural equation modeling with near singular covariance matrices. *Computational Statistics & Data Analysis, 52,* 4842–4858.

Yuan, K.-H., & Chan, W. (2016). Structural equation modeling with unknown population distributions: Ridge generalized least squares. *Structural Equation Modeling, 23,* 163–179.

Equivalence Testing for Factor Invariance Assessment with Categorical Indicators

W. Holmes Finch and Brian F. French

Abstract Factorial invariance assessment is central in the development of educational and psychological instruments. Establishing factor structure invariance is key for building a strong validity argument, and establishing the fairness of score use. Fit indices and guidelines for judging a lack of invariance is an ever-developing line of research. An equivalence testing approach to invariance assessment, based on the RMSEA has been introduced. Simulation work demonstrated that this technique is effective for identifying loading and intercept noninvariance under a variety of conditions, when indicator variables are continuous and normally distributed. However, in many applications indicators are categorical (e.g., ordinal items). Equivalence testing based on the RMSEA must be adjusted to account for the presence of ordinal data to ensure accuracy of the procedures. The purpose of this simulation study is to investigate the performance of three alternatives for making such adjustments, based on work by Yuan and Bentler (Sociological Methodology, 30(1):165–200, 2000) and Maydeu-Olivares and Joe (Psychometrika 71(4):713–732, 2006). Equivalence testing procedures based on RMSEA using this adjustment is investigated, and compared with the Chi-square difference test. Manipulated factors include sample size, magnitude of noninvariance, proportion of noninvariant indicators, model parameter (loading or intercept), and number of indicators, and the outcomes of interest were Type I error and power rates. Results demonstrated that the T_3 statistic (Asparouhov & Muthén, 2010) in conjunction with diagonally weighted least squares estimation yielded the most accurate invariance testing outcome.

Keywords Invariance testing · Equivalence test · Categorical indicator

W. Holmes Finch (✉)
Ball State University, Muncie, IN 47306, USA
e-mail: whfinch@bsu.edu

B. F. French (✉)
Washington State University, Pullman, WA 99164, USA

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_21

1 Introduction

Social scientists, policy makers, and others make use of scores from psychological scales to make decisions about persons, and groups of people, for a variety of purposes, including hiring, school matriculation, professional licensure, and determinations regarding the need for special educational and psychological services. Given their importance, there must be strong validity evidence for using scores in these ways (American Education Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). One important aspect of providing such evidence is the determination as to whether the measures provide equivalent information for members of different groups in the population, such as males and females, or members of different economic subgroups (Wu, Li, & Zumbo, 2007). Traditionally, such factor invariance (FI) assessments have been made using a Chi-square difference test with multiple group confirmatory analysis (MGCFA). However, this approach is very sensitive to sample size, so that it might be statistically significant for very minor differences in group parameter values (Yuan & Chan, 2016). Perhaps more importantly, information about the magnitude of any group differences in latent variable model parameters identified is not available (Yuan & Chan, 2016). Yuan and Chan described an alternative approach to FI assessment that is based on equivalence testing. When indicator variables are normally distributed, this equivalence testing based method is an effective tool, yielding accurate results with respect to the invariance (or noninvariance) of the latent variable model (Finch & French, 2018). The purpose of the current simulation study was to extend this earlier work by investigating how the equivalence testing technique performed when the observed indicators were ordinal variables (such as items on a scale), rather than being normally distributed.

1.1 MGCFA and FI Assessment

FI assessment (Millsap, 2011) refers to a set of nested models with differing levels of cross group equality assumed about the parameters in a latent variable model linking observed indicators (x) to latent variables (ξ). The weakest type of FI is configural invariance (CI), where only the general latent structure (i.e., number of latent variables and correspondence of observed indicators to latent variables) is the same across groups. The next level of FI is measurement invariance (MI), where the factor loading matrix (Λ) is assumed to be equivalent across groups (Kline, 2016; Wicherts & Dolan, 2010). If MI holds, researchers might next assess the equality of the factor model intercepts (τ) across groups (Steenkamp & Baumgartner, 1998), and/or group equality of the unique variances (δ) invariant across groups.

The most common approach for assessing FI is based on the MGCFA model:

$$x_g = \tau_g + \Lambda_g \xi + \delta_g \tag{1}$$

where

x_g Observed indicators for group g
τ_g Threshold parameters for group g
Λ_g Factor loading matrix for group g
ξ Latent variable(s)
δ_g Unique variances of the indicator variables for group g.

The terms in Eq. (1) are as described above, except that the parameters are allowed to vary by group, which is denoted by the g subscript. MGCFA is used to test each type of FI through a series of nested models, which differ in terms of that model parameters that are held equivalent between groups. For example, in order to assess MI, the researcher would constrain factor loadings to be equivalent across groups, thereby replacing Λ_g with Λ in Eq. (1). The fit of the constrained and unconstrained models are then compared using a difference in Chi-square fit statistic value, χ^2_Δ. The null hypothesis of this test is that MI is present.

The performance of χ^2_Δ for invariance testing has yielded somewhat mixed results. French and Finch (2006) found that for normally distributed indicators and a sample size of no more than 500, χ^2_Δ had Type I error rates at the nominal (0.05) level, while also exhibiting relatively high power. Other researchers have reported that χ^2_Δ is sensitive to sample size, to a lack of normality in the indicators, and to model misspecification errors, and in such cases may yield inflated Type I error rates when assessing MI (Chen, 2007; Yuan & Chan, 2016; Yuan & Bentler, 2004).

1.2 Factor Invariance with Equivalence Testing

Given these problems associated with using χ^2_Δ, Yuan and Chan (2016) proposed an extension of other work designed to assess model fit using an equivalence testing approach (Marcoulides & Yuan, 2017; Yuan, Chan, Marcoulides, & Bentler, 2016) to the assessment of FI using MGCFA. In the case of MGCFA for FI assessment, the null hypothesis is:

$$H_{0I} : (F_{bc0} - F_{b0}) > \varepsilon_0 \tag{2}$$

F_{bc0} is the fit function value for a model where group parameters are constrained to be equal, F_{b0} is the fit function value for a model where group latent variable model parameters are allowed to differ between groups, and ε_0 is the maximum acceptable model misspecification. Rejecting H_{0I} leads to the conclusion that any model misspecification due to constraining factor model parameters to be equal across groups does not greatly degrade model fit vis-à-vis the model where these constraints are relaxed. Therefore, rejecting H_{0I} in the MI equivalence testing framework would indicate that when the groups' factor loadings are constrained to be equal, the difference in fit between the loadings constrained and loadings unconstrained models does not exceed an acceptable level of misfit, as expressed by ε_0.

Yuan and Chan (2016) showed that the value of ε_0 can be obtained as follows:

$$\varepsilon_0 = \frac{df(RMSEA_0)^2}{m} \tag{3}$$

where

df	Model degrees of freedom
m	Number of groups
$RMSEA_0$	Maximum value of *RMSEA* that can be tolerated.

For FI assessment, Yuan and Chan recommend using this equivalence testing procedure to characterize the relative degree of noninvariance present in the data, as opposed to making strict hypothesis testing based determinations regarding equivalence or not. In this framework, the degree of model parameter invariance present in the data can be characterized using common guidelines (e.g., MacCallum, Browne, & Sugawara, 1996) to describe the model constraining group parameters to be equal. These guidelines for interpreting values of RMSEA suggest the following fit categories: Excellent fit (<0.01), Close fit (0.01–0.05), Fair fit (0.05–0.08), Mediocre fit (0.08–0.10), and Poor fit (0.10+). Thus, an RMSEA of 0.17 for a model constraining factor loadings to be equal among groups would suggest poor fit of the MI model, meaning that model parameters are likely not equivalent between the groups. Yuan and Chan (2016) found that for the purposes of determining the value of ε_0, these standard cutoffs for interpreting *RMSEA* may be too stringent, and thus recommended an alternative approach for obtaining adjusted cutoffs based on the data being analyzed. The interested reader is encouraged to review this earlier paper for a description of how these alternatives are obtained. This equivalence testing approach is effective for assessing the fit of a single model, and for invariance assessment (e.g., Finch & French, 2018; Marcoulides & Yuan, 2017; Yuan & Chan, 2016). However, the performance of the equivalence testing approach to invariance assessment when indicators are categorical and not normally distributed has not been investigated.

1.3 Fit Indices for Categorical Indicators

Yuan and Chan (2016) indicated that the equivalence invariance test was designed for use with normally distributed indicators. However, in many contexts in the social sciences researchers work with ordinal observed variables, such as responses to items on a rating scale. In such cases, the equivalence testing approach may not be appropriate, because calculation of the standard full information χ^2 statistic upon which RMSEA is based is problematic (Maydeu-Olivares & Joe, 2006). In the context of categorical indicators, this statistic relies on the full cross-tabulation of the entire set of categorical indicators (full information), leading to the potential for cell sparsity, and resulting problems in its calculation (Maydeu-Olivares & Joe), which in turn biases the RMSEA estimate.

In order to address these problems caused by sparsity, alternative goodness of fit statistics based on limited information approaches have been proposed for use with latent variable modeling in the context of categorical indicator variables. One set of alternatives is based upon a least squares, rather than maximum likelihood, estimation paradigm. For example, the weighted least squares (WLS) fit function takes the form:

$$F(\theta;W) = \big(\hat{\rho} - \rho(\theta)\big)' W \big(\hat{\rho} - \rho(\theta)\big) \tag{4}$$

where

$\hat{\rho}$	Sample polychoric correlation matrix for the indicator variables
$\rho(\theta)$	Model implied polychoric correlation matrix
W	Asymptotic covariance matrix of $\hat{\rho}$.

Given that WLS yields biased estimates and has difficulty in converging when samples are small (Muthén, 1993), the diagonally weighted least squares (DWLS) estimator was proposed (Muthén, du Toit, & Spisic, 1997). DWLS reduces the computational burden and yields less biased parameter estimates for smaller sample sizes by using only the diagonal of W as the weight matrix (Flora & Curran, 2004). When W is the identity matrix, (5) is the unweighted least squares (ULS) estimator. For each of these estimators, a moment corrected goodness of fit statistic, T_3, can be calculated based upon the fit function, and is asymptotically a Chi-square statistic (Asparouhov & Muthén, 2010). T_3 can then be used to calculate RMSEA, which in turn can be used with the invariance equivalence methodology described above.

An alternative limited information goodness of fit statistic for use with categorical indicators was proposed by Maydeu-Olivares and Joe (2006). This statistic is defined as:

$$M_2^* = N\hat{e}_2\hat{\Omega}_2\hat{e}_2 \tag{5}$$

where

$\hat{e}_2$	Vector of first and second order residual probabilities.
$\hat{\Omega}_2$	$\hat{\Omega}_2 = \Xi_2^{-1} - \Xi_2^{-1}\Delta_2\big(\Delta_2'\Xi_2^{-1}\Delta_2\big)^{-1}\Delta_2'\Xi_2^{-1}$
Ξ_2	Asymptotic covariance matrix of the first and second order sample proportions
Δ_2	Matrix of derivatives of the first and second order model implied probabilities with respect to the vector of parameter estimates $\hat{\theta}$.
M_2^*	is asymptotically distributed as a Chi-square statistic, and can be used to calculate RMSEA for use with the invariance equivalence testing approach described above.

1.4 Goals of the Current Study

The goal of the current study was to extend earlier work that investigated the performance of the equivalence testing procedure for normally distributed indicators (Finch & French, 2018). The current study extends this research by examining the performance of T_3 for both DWLS and ULS, as well as M_2 in the context of MI when the indicator variables are categorical.

2 Method

A Monte Carlo simulation study (1000 replications) was utilized to address the study goals. Data simulation was completed in Mplus, version 7.11 (Muthén & Muthén, 1998–2016), and data analyses were conducted using R version 3.3.1 (R Development Core Team, 2016). Data were generated using a single factor confirmatory factor analysis model for 2 groups, where the factor, error variances, and factor variances followed the standard normal distribution, with a mean of 0 and variance of 1. Indicator variables were simulated to be ordinal with 5 categories, with the following pattern of thresholds: $-1, -0.5, 0.5, 1$. Factor loadings were set to 1 for all indicators, unless manipulated to induce measurement noninvariance, as described below. All other model parameters were held invariant between the two groups. The referent indicator method was used to identify the factor models. The following factors were manipulated in the study, and were based upon earlier published work in this area (e.g., Finch & French, 2018).

2.1 Sample Size

Given that sample size has been shown to be important in terms of the performance of the equivalence testing approach, and the χ^2_Δ test (Chen, 2007; Finch & French, 2018; French & Finch, 2006), it was manipulated in the current study. Total sample sizes were simulated to be 200, 400, 600, 1000, 1500, or 2000, and were designed to reflect small to large samples.

2.2 Number of Indicator Variables

Either 10 or 20 observed factor indicators were simulated, representing a range of values that might be encountered in practice.

2.3 Number of Noninvariant Indicators and Magnitude of Measurement Noninvariance

Measurement noninvariance was simulated by creating group differences in the factor loadings for some observed indicators. For the invariance condition, the difference in factor loadings between the groups was 0 (complete invariance). For the noninvariant cases, loadings were simulated to differ by 0.1, 0.2, 0.3, 0.4, or 0.5. The percent of indicators allowed to be noninvariant was 0, 10, 20, or 30%. As an example of how noninvariance was simulated, in the 10 indicators, 10% noninvariant, 0.1 noninvariance magnitude condition, the factor loading for indicator 2 was set to 0.9 in one group, and kept at 1.0 in the other group.

2.4 Invariance Assessment Approaches

For each replication within each simulation condition, invariance was tested using the MGCFA χ^2_Δ approach, with T_3 for DWLS (T_{DWLS}) and ULS (T_{ULS}), as well as M_2^*. In addition, the equivalence test method based was also used to assess invariance, with the RMSEA values based upon T_{DWLS}, T_{ULS}, and M_2^*, respectively.

2.5 Study Outcomes

The outcomes were the Type I error and power rates of the χ^2_Δ tests, and the adjusted equivalence test fit category distribution (Excellent, Close, Fair, Mediocre, or Poor). Analysis of variance (ANOVA) was used to identify statistically significant main effects and interactions of the manipulated conditions with respect to the proportion of cases for which the equivalence testing method identified poor fit. In addition, the partial η^2 effect size was also used to identify ANOVA model terms of interest, such that main effects and interactions of the manipulated conditions had to be statistically significant with partial η^2 value of 0.1 or larger, ensuring that effects accounted for at least 10% of the outcome variance to be deemed important.

3 Results

3.1 Measurement Invariance Is Present

The interaction of invariance assessment method by sample size was the only statistically significant model term ($F_{10,8} = 10.527$, $p = 0.001$, $\eta^2 = 0.929$) when invariance was present. The Type I error rate for the T_{DWLS} statistic was the only one

Table 1 Type I error rates and proportion of adjusted equivalence test results in excellent or close range, by sample size and method

N	T_{DWLS}	T_{ULS}	M_2^*	Proportion in excellent/close fit range		
				RMSEA T_{DWLS}	RMSEA T_{ULS}	RMSEA M_2^*
200	0.07	0.08	0.11	0.39	0.12	0.03
400	0.06	0.08	0.10	0.41	0.13	0.03
600	0.04	0.08	0.10	0.57	0.21	0.04
1000	0.05	0.08	0.07	0.80	0.45	0.25
1500	0.05	0.07	0.06	0.84	0.55	0.34
2000	0.05	0.07	0.05	0.92	0.69	0.50

that was in the acceptable range (0.025–0.075) as defined by Bradley (1978), across all sample size conditions (Table 1). For the other two statistics, the samples needed to be at least 1000 (M_2^*) or 1500 (T_{ULS}) in order for the Type I error rates to be in this range. Results in Table 1 show that equivalence testing based on T_{DWLS} had the highest rates in the expected excellent/close fit categories, across sample sizes. This proportion increased concomitantly with increases in sample size. Finally, the proportion in the expected excellent/close fit range was below 0.8 for samples of less than 1000 for T_{DWLS}, which was the best performer in this regard.

3.2 Measurement Invariance Is Not Present

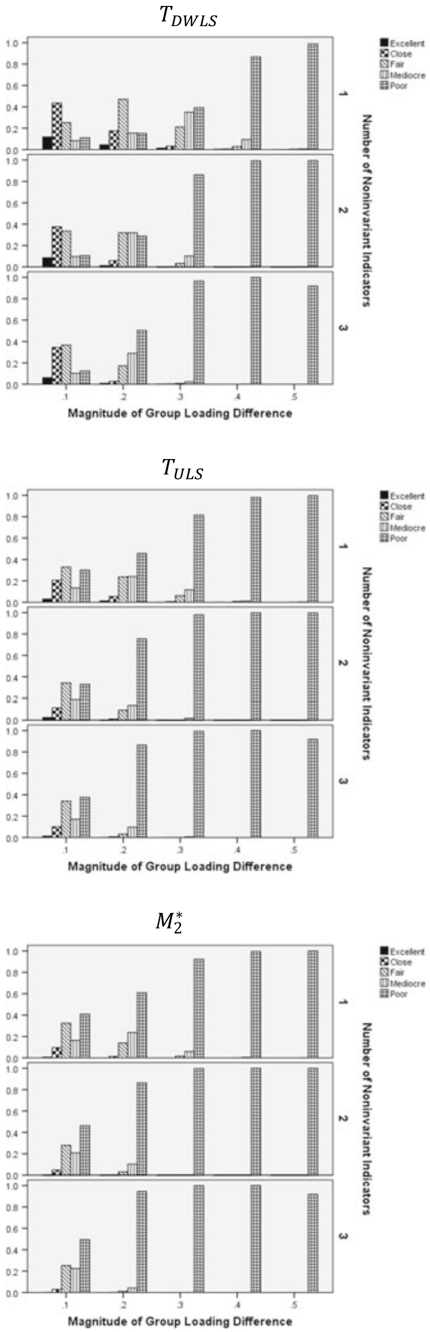
When factor loadings were simulated to differ between the groups, ANOVA found the interactions of equivalence test statistic by number of noninvariant indicator variables by magnitude of group loading difference ($F_{16,234} = 14.389$, $p < 0.001$, $\eta^2 = 0.496$), and equivalence test statistic by number of loadings by magnitude of group loading difference ($F_{8,234} = 9.104$, $p < 0.001$, $\eta^2 = 0.237$), to be statistically significantly related to the performance of the equivalence test procedure. The first set of results to be examined are those for measurement invariance not present, by method, magnitude of group loading difference, and percent of noninvariant loadings. The proportion of replications in each equivalence testing category for this combination of conditions appear in Fig. 1. It is clear from these results that when the magnitude of group loading differences was 0.3 or more, and 20 or 30% of the indicators were noninvariant between groups, virtually all replications were in the poor fit range (expected outcome given simulated lack of invariance) for all of the methods. Under conditions in which the degree of group difference was less pronounced, the invariance tests based on M_2^* and T_{ULS} tended to indicate worse fit more frequently than did those based on T_{DWLS}. This result was strongest when 30% of the indicators were simulated to have different loadings between groups, and the magnitude of these differences was 0.1 or 0.2. Power results for the χ^2_Δ tests appear

Table 2 Power rates for detecting measurement noninvariance for the chi-square difference test by the magnitude of group loading difference, number of noninvariant indicators, and test statistic

Magnitude of group loading difference	Number of noninvariant indicators	T_{DWLS}	T_{ULS}	M_2^*
0.1	1	0.07	0.84	0.15
	2	0.11	0.90	0.28
	3	0.15	0.90	0.29
0.2	1	0.21	0.93	0.44
	2	0.45	0.97	0.63
	3	0.61	0.98	0.67
0.3	1	0.53	0.97	0.63
	2	0.76	0.99	0.82
	3	0.85	0.99	0.87
0.4	1	0.76	0.99	0.79
	2	0.91	0.99	0.92
	3	0.97	1.00	0.95
0.5	1	0.89	0.99	0.89
	2	0.94	0.99	0.94
	3	0.99	1.00	0.990

in Table 2, and demonstrate that T_{ULS} had the highest rates of power across conditions, whereas T_{DWLS} exhibited somewhat lower power than did M_2^*, particularly for lower group loading difference magnitudes, and fewer noninvariant indicators. It is important when interpreting these results to recall that the Type I error rates were inflated under many conditions for each of these statistics, particularly M_2^* and T_{ULS}.

The proportion of replications in each equivalence testing category by magnitude of group loading difference and number of indicators when noninvariance was simulated to be present appear in Fig. 2. These results revealed that with a larger group loading difference there was a higher likelihood of mediocre and poor fit, based on the equivalence test. In addition, with more indicators this effect was magnified for each of the statistics. For example, the proportion of cases in the mediocre and poor fit categories was greater for 20 indicators than for 10, across methods studied here. Power results for the χ^2_Δ tests by magnitude of group loading difference and number of indicators appear in Table 3, and are aggregated over the number of non-invariant indicators. Power for all three equivalence testing methods was higher when more indicators were present, and that power for T_{ULS} was the highest across conditions, whereas power for T_{DWLS} was the lowest for the smallest magnitudes of group loading difference.

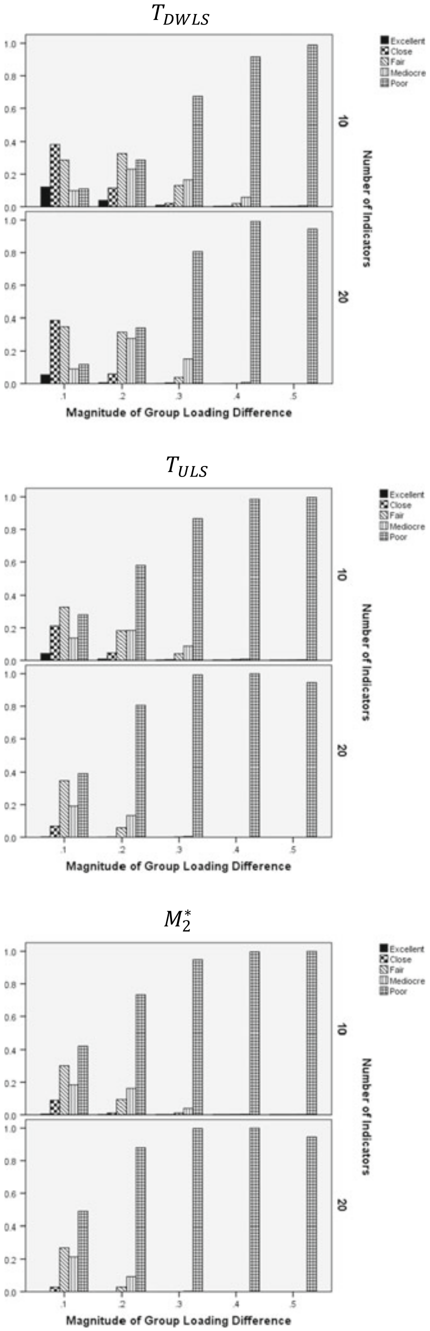


T_{DWLS}

T_{ULS}

M_2^*

Fig. 1 Proportion of adjusted equivalence test results in each fit category by equivalence statistic, number of noninvariant loadings, and magnitude of group loading difference: noninvariance present



T_{DWLS}

T_{ULS}

M_2^*

Fig. 2 Proportion of adjusted equivalence test results in each fit category by equivalence statistic, number of indicator variables, and magnitude of group loading difference: noninnvariance present

Table 3 Power rates for detecting measurement noninvariance for the chi-square difference test by the magnitude of group loading difference, number of indicators, and test statistic

Magnitude of group loading difference	Number of indicators	T_{DWLS}	T_{ULS}	M_2^*
0.1	10	0.10	0.79	0.30
	20	0.11	0.97	0.39
0.2	10	0.38	0.93	0.58
	20	0.46	0.99	0.68
0.3	10	0.66	0.98	0.74
	20	0.77	0.99	0.81
0.4	10	0.84	0.99	0.85
	20	0.92	1.00	0.92
0.5	10	0.91	0.99	0.91
	20	0.97	1.00	0.97

4 Discussion

The results of this study demonstrated that the equivalence testing procedure based on T_{DWLS} appeared to correctly identify models in which MI held at the highest rates among the methods studied here, while at the same time generally identifying poorly fitting models at a high rate. It is important to note that when the magnitude of group factor loading difference was relatively low (0.2 or less), this statistic was less likely to indicate fair to poor fit than the alternatives studied here. This result could suggest a relative lack of power for this approach, or it could simply reflect the fact that small differences in factor loadings are not indicative of a major lack of equivalence between groups. Finally, the χ^2_Δ based approaches exhibited inflated Type I error rates in many cases, and may not be as useful as the equivalence testing approach.

Future research in this area should focus on identifying additional alternatives for calculating RMSEA with categorical indicators. Though T_{DWLS} was the best performer, it was not without problems, particularly for low levels of noninvariance. In addition, future work should include a wider array of indicator categories (e.g., 3, 4, 6, 7), and more complex latent structure (e.g., 2 or 3 factors). Such continued work will allow the invariance literature to continue to expand to address group differences in the measurement of constructs used to make decisions about individuals.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational & Psychological Testing*. Washington, D.C.: American Educational Research Association.

Asparouhov, T., & Muthén, B. O. (2010). *Simple second order chi-square correction*. Retrieved from: http://www.statmodel.com/download/WLSMV_new_chi21.pdf.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31,* 321–339.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504.

Finch, W. H., & French, B. F. (2018). A simulation investigation of the performance of invariance assessment using equivalence testing procedures. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(5), 673–686.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466–491.

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 13*, 378–402.

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. New York: The Guilford Press.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130–149.

Marcoulides, K. M., & Yuan, K.-H. (2017). New ways to evaluate goodness of fit: A note on using equivalence testing to assess structural equation models. *Structural Equation Modeling, 24*(1), 148–153.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*(4), 713–732.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Muthén, B. (1993). Goodness of fit with categorical and other nonnormal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–234). Newbury Park, CA: Sage.

Muthén, L. K., & Muthén, B. O. (1998–2016). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, B. O., du Toit, S. H., Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished technical report.

R Development Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Steenkamp, J-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*(1), 78–90.

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement Issues and Practice, 29*(3), 39–47.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMMS data. *Practical Assessment, Research & Evaluation, 12*(3), 1–26.

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology, 30*(1), 165–200.

Yuan, K.-H., & Bentler, P. M. (2004). On chi-square difference and Z tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement, 64,* 737–757.

Yuan, K.-H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square difference tests. *Psychological Methods, 21*(3), 405–426.

Yuan, K.-H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(3), 319–330.

Canonical Correlation Analysis with Missing Values: A Structural Equation Modeling Approach

Zhenqiu (Laura) Lu

Abstract Canonical correlation analysis (CCA) is a generalization of multiple correlation that examines the relationship between two sets of variables. When there are missing values, spectral decomposition in CCA becomes complicated and difficult to implement. This article investigates structural equation modeling approach to Canonical correlation analysis when data have missing values.

Keywords Canonical correlation analysis · Structural equation modeling · Missing values

1 Introduction

Canonical correlation analysis (CCA) (see Hotelling, 1936) is a statistical method to calculate multiple correlations that examines the relationship between two sets of variables. In a stepwise procedure, pairs of linear combinations of original variables, one from each set, are derived. At each step, the linear combinations are derived to maximize the correlation between them. Also, the current pair of combinations is uncorrelated with all previously derived pairs (Anderson, 2003). The number of such pairs is the number of variables in the smaller set of original variables. The series of maximal correlations between pairs of newly created combinations are called *canonical correlations* (CCs). 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 canonical correlations are exclusively determined by the canonical weight coefficients. For normalization purposes, the weight coefficients must satisfy some restrictions.

The goal of CCA is essentially to find the optimal weights that maximize these canonical correlations. However, the actual implementations of CCA are not computationally effective. The traditional way is to employ a mathematically equivalent

Z. (Laura) Lu (✉)
University of Georgia, 325V Aderhold Hall, 110 Carlton Street, Athens, GA 30602, USA
e-mail: zlu@uga.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_22

spectral decomposition of some quadruple product of covariance/correlation matrices, and all canonical correlations and the associated weight coefficients are obtained simultaneously. There are many limitations for this approach. For example, it is hard to deal with missing values. When there are missing values at any set of original variables, the spectral decomposition in CCA becomes really complicated and difficult to implement. The traditional way is to conduct a list-wise deletion to make data complete.

Structural equation modeling (e.g., Bollen, 1989) is widely used statistical method to investigate the underlying relationship among observed variables and latent variables. It is a generalization of various multivariate linear models, such as path analysis, measurement models, factor models, structural relation models, and latent growth models. A general SEM model has three parts, one measurement model of exogenous variables, one measurement model of endogenous variables, and an overarching structural model of the relationships among exogenous and endogenous variables.

The research on statistical connections between CCA and SEM is less obvious and is rare in literature. Bagozzi, Fomell, and Larcker (1981) discussed that canonical correlation analysis could be viewed as a case of a structural relations model. Following this idea, Fan (1997) developed a multiple stages procedure using a multiple indicators and multiple causes (MIMIC) model at each step to analyze canonical correlations. But "the representation of CCA using SEM is not straightforward" (Fan, 1997, p. 69). Also, as it is a multiple stages approach, it is not practical to conduct the stepwise procedure for most researchers. Lu and Gu (2018) proposed a general SEM approach to CCA. The CCA is directly mathematically formulated by the structural equation modeling. However, all the researchers above did not investigate the SEM approach to CCA with missing values.

Missing data are almost inevitable. Research participants may drop out of a study, or some students may miss a test due to absence or fatigue. Missing data can be investigated from their mechanisms, or why missing data occur. Little and Rubin distinguished *ignorable missingness mechanism* and *non-ignorable missingness mechanism*. For ignorable missingness mechanism, estimates are usually asymptotically consistent when the missingness is ignored, because the parameters that govern the missing process either are distinct from the parameters that govern the model outcomes or depend on the observed variables in the model. The ignorable missingness mechanism includes *missing completely at random* (MCAR), in which the missing data probability does not depend either on observed outcomes or on missing values (or latent variables), and *missing at random* (MAR), in which the missing data probability may depend on some observed outcomes, but not on missing values (or latent variables).

By using SEM, missing data can be easily handled with maximum likelihood (ML) estimation method under the MCAR or MAR assumption, and this method is now available in most popular SEM software.

In this article, we investigate the structural equation modeling approach to canonical correlation analysis when the data have missing values. We compare the results obtained from the conventional CCA approach and those from the SEM approach by using maximum likelihood estimation method.

2 SEM Approach to CCA

In this section, we describe the structural equation modeling approach to canonical correlation analysis.

2.1 SEM Representation of 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 $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ are the covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively, and $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}'_{21}$ is the covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$. So if we use $\mathbf{Z} = \left[\mathbf{X}' \ \mathbf{Y}'\right]'$, then the covariance matrix of $\mathbf{Z}$ is

$$Cov(\mathbf{Z}) = \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

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 $E(V_{1i}V_{2i}) = E\left(\mathbf{a}'_{1i}\mathbf{X}\mathbf{Y}'\mathbf{a}_{2i}\right) = \mathbf{a}'_{1i}\boldsymbol{\Sigma}_{12}\mathbf{a}_{2i}$ ($i = 1, 2, \ldots, p$), subject to the following unit-variance and orthogonality constraints as follows, unit-variance $\mathbf{a}'_{1i}\boldsymbol{\Sigma}_{11}\mathbf{a}_{1i} = \mathbf{a}'_{2i}\boldsymbol{\Sigma}_{22}\mathbf{a}_{2i} = 1$, ($i = 1, 2, \ldots, p$), within-set orthogonality $\mathbf{a}'_{1i}\boldsymbol{\Sigma}_{11}\mathbf{a}_{1j} = \mathbf{a}'_{2i}\boldsymbol{\Sigma}_{22}\mathbf{a}_{2j} = 0$, ($i \neq j$ and $i, j = 1, 2, \ldots, p$) and between-set orthogonality $\mathbf{a}'_{1i}\boldsymbol{\Sigma}_{12}\mathbf{a}_{2j} = 0$, ($i \neq j$ and $i, j = 1, 2, \ldots, p$). If we let $\mathbf{A}_1 = \left(\mathbf{a}_{11} \ldots \mathbf{a}_{1p}\right)$, and $\mathbf{A}_2 = \left(\mathbf{a}_{21} \ldots \mathbf{a}_{2p}\right)$. By adding another $(p + d) \times d$ additional matrix $\mathbf{A}_3 = \left(\mathbf{a}_{2,p+1} \ldots \mathbf{a}_{2,p+d}\right)$ with each column derived from variable $\mathbf{Y}_2$ (Anderson, 2003, p. 499). We further assume the unit-variance and orthogonality constraints for matrix $\mathbf{A}_3$. Let $\mathbf{A}$ be a $(2p + d) \times (2p + d)$ block-diagonal matrix

$$\mathbf{A} = \left(\begin{array}{c|cc} \mathbf{A}_1 & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{A}_2 & \mathbf{A}_3 \end{array}\right),$$

where $\mathbf{A}_1$ 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 definition and constraints above, the conventional CCA states that mathematically

$$\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A} = \begin{pmatrix} \mathbf{A}'_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}'_2 \\ \mathbf{0} & \mathbf{A}'_3 \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{A}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{A}_3 \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{A'}_1\mathbf{\Sigma}_{11}\mathbf{A}_1 & \mathbf{A'}_1\mathbf{\Sigma}_{12}\mathbf{A}_2 & \mathbf{A'}_1\mathbf{\Sigma}_{12}\mathbf{A}_3 \\ \mathbf{A'}_2\mathbf{\Sigma}_{21}\mathbf{A}_1 & \mathbf{A'}_2\mathbf{\Sigma}_{22}\mathbf{A}_2 & \mathbf{A'}_2\mathbf{\Sigma}_{22}\mathbf{A}_3 \\ \mathbf{A'}_3\mathbf{\Sigma}_{21}\mathbf{A}_1 & \mathbf{A'}_3\mathbf{\Sigma}_{22}\mathbf{A}_2 & \mathbf{A'}_3\mathbf{\Sigma}_{22}\mathbf{A}_3 \end{pmatrix} = \left(\begin{array}{c|c|c} \mathbf{I}_p & \mathbf{R} & \mathbf{0} \\ \hline \mathbf{R} & \mathbf{I}_p & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{I}_d \end{array}\right)$$

where **0** denotes a matrix of zeros of proper dimension. In other words, if we assume the vectors of canonical variates $\mathbf{V}_1 = \left(V_{11} \cdots V_{1p}\right)'$ and $\mathbf{V}_2 = \left(V_{21} \cdots V_{p+d}\right)'$, then

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{bmatrix} = \mathbf{A}'Z = \mathbf{A}' \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \begin{pmatrix} \mathbf{A'}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{A'}_2 \\ \mathbf{0} & \mathbf{A'}_3 \end{pmatrix} \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \tag{1}$$

with a covariance matrix

$$Cov(\mathbf{V}) = \mathbf{A}'\mathbf{\Sigma}\mathbf{A} = \left(\begin{array}{c|c|c} \mathbf{I}_p & \mathbf{R} & \mathbf{0} \\ \hline \mathbf{R} & \mathbf{I}_p & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{I}_d \end{array}\right) \tag{2}$$

Equation (1) can be transformed to an equivalent form

$$\mathbf{Z} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \left(\mathbf{A}'\right)^{-1}\mathbf{V} = \left(\mathbf{A}'\right)^{-1} \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{bmatrix} \tag{3}$$

with Cov(**V**) as in (2). Equation (3) is a simplified measurement model in SEM when $\mathbf{\Lambda}_{\mathrm{y}} = (\mathbf{A}')^{-1}$, $\boldsymbol{\eta} = \mathbf{V}$, and $\boldsymbol{\varepsilon} = \mathbf{0}$. Therefore, the matrix form of CCA has been represented by a simplified structural equation model,

$$\mathbf{Z} = \left(\mathbf{A}'\right)^{-1}\mathbf{V} + \mathbf{0} = \mathbf{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon},$$

where $\mathbf{\Lambda}$ is a factor loading matrix for the vector of observed indicators **Z**, $\boldsymbol{\eta}$ is a vector of latent variables, $\boldsymbol{\varepsilon}$ is a vector of measurement errors of **Z**. Because of $\mathbf{A}' = \mathbf{\Lambda}^{-1}$, the weight matrix **A** in CCA can be obtained as the transpose of the inverse of the loading matrix in SEM.

2.2 SEM Approach to CCA with Missing Values

Based on the derivation above, the CCA has been represented by a simplified structural equation model. To address the missing data problem in SEM, researchers have developed many approaches (Enders & Bandalos, 2001; Yuan & Lu, 2008). The most widely used method is the full information maximum likelihood (ML) estimation method.

For complete data, the log likelihood function for independent observations from a distribution with a probability density function f (.) can be expressed as

$$l = \log(L) = \sum_{i=1}^{N} \log(f(\mathbf{x}_i|\boldsymbol{\theta}))$$

where $\mathbf{x}_\mathrm{i}$ is a vector of observed values for individual *i*, $\boldsymbol{\theta}$ is a vector of parameters, such as mean vector and covariance matrix for multivariate normal distribution. When data have missing values, the likelihood becomes

$$l = \sum_{i=1}^{N} \log(f(\mathbf{x}_i|\boldsymbol{\theta}_i))$$

where the forms of $\mathbf{x}_\mathrm{i}$ and $\boldsymbol{\theta}_\mathrm{i}$ depend on individual *i*. If there are missing values for that individual, then $\mathbf{x}_\mathrm{i}$ is a sub-vector of the complete case $\mathbf{x}_\mathrm{i}$ by deleting corresponding missing elements, and $\boldsymbol{\theta}_\mathrm{i}$ is sub-matrix of $\boldsymbol{\theta}$ by deleting the rows and column corresponding to missing elements. The likelihood function with missing values is then maximized to obtain ML estimates. The maximization can be carried out by conventional numerical methods, such as the Newton-Raphson algorithm.

2.3 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.

The SEM approach to CCA can be implemented by using existing SEM software packages, such as the Lavaan package (Rosseel, Oberski, Byrnes, Vanbrabant, & Savalei, 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 use the R package Lavaan, Mplus, and EQS. The results from these SEM software are similar. Special settings or options in SEM might be included, such as the option of "missing = ml" in Lavann in order to use the full information maximum likelihood (FIML) estimation by using all available data when data are missing completely at random (MCAR) or missing at random (MAR).

3 Real Data Analysis

In this section, we illustrate the SEM approach to CCA with missing values.

3.1 The Data Set

The data are from 34 countries of the world from FAOSTAT (Food and Agriculture Organization of the United Nations, 1998). We are interested in two sets of variables: one on food supply and the other on cancer type. By using canonical correlation analysis, we are trying to investigate the relationships between food intake types and the mortality rate by cancer types.

In the data set there are four cancer sites, esophagus, stomach, pancreas, and liver. Suppose we use a 4 by 1 vector X = (x1, x2, x3, x4)′ to represent these variables. Also, the data include seven food supplies, alcohol, meat, fish, cereal, vegetable, milk products, and the total calorie per day. We use a 7 by 1 vector Y = (y1, y2, y3, y4, y5, y6, y7)′ to represent these food supplies. In total, there are 34 complete cases, with 11 variables for each case. The total observed values are $34 \times 11 = 374$.

3.2 Generating Incomplete Data Set with Missing Values

In order to generate data sets with missing values, we randomly deleted some observed values from the complete data to make the missing mechanism MCAR. We first created an incomplete data set with a low missing data rate by removing only 2 observations. The missing data rate is $2/(34 * 11) = 0.53\%$. But if it's a listwise deletion, then two rows will be removed and the missing data rate becomes $(2 * 11)/(34 * 11) = 5.88\%$.

We also created a second MCAR incomplete data set by removing more observations. The missing data rate is $15/(34 * 11) = 4.01\%$. Again, if the list-wise deletion is conducted, then the five rows will be removed, and the missing data rate becomes $(5 * 11)/(34 * 11) = 14.71\%$.

3.3 SEM Model

The factor model can be set up as follows.

$$\begin{bmatrix}\mathbf{X}\\ \mathbf{Y}\end{bmatrix} = \left[\begin{array}{c} x_1\\ x_2\\ x_3\\ x_4\\ \hline y_1\\ y_2\\ y_3\\ y_4\\ y_5\\ y_6\\ y_7 \end{array}\right] = \boldsymbol{\Lambda\eta} + \mathbf{0} = \left(\begin{array}{cccc|ccccccc} \lambda_{11} & & sym & & & & & & & & \\ \lambda_{21} & \lambda_{22} & & & & & & 0 & & & \\ \lambda_{31} & \lambda_{32} & \lambda_{33} & & & & & & & & \\ \lambda_{41} & \lambda_{42} & \lambda_{43} & \lambda_{44} & & & & & & & \\ \hline & & & & \lambda_{55} & & & & & & \\ & & & & \lambda_{65} & \lambda_{66} & & & sym & & \\ & & & & \lambda_{75} & \lambda_{76} & \lambda_{77} & & & & \\ & & & & \lambda_{85} & \lambda_{86} & \lambda_{87} & \lambda_{88} & & & \\ & & 0 & & \lambda_{95} & \lambda_{96} & \lambda_{97} & \lambda_{98} & \lambda_{99} & & \\ & & & & \lambda_{105} & \lambda_{106} & \lambda_{107} & \lambda_{108} & \lambda_{109} & \lambda_{1010} & \\ & & & & \lambda_{115} & \lambda_{116} & \lambda_{117} & \lambda_{118} & \lambda_{119} & \lambda_{1110} & \lambda_{1111} \end{array}\right) \left[\begin{array}{c} \eta_1\\ \eta_2\\ \eta_3\\ \eta_4\\ \hline \eta_5\\ \eta_6\\ \eta_7\\ \eta_8\\ \eta_9\\ \eta_{10}\\ \eta_{11} \end{array}\right]$$

$$= \left(\begin{array}{c|c} \mathbf{\Lambda}_1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{\Lambda}_2 \end{array}\right)\boldsymbol{\eta} = (\mathbf{A}')^{-1}\boldsymbol{\eta}$$

where $\boldsymbol{\eta}$ having a covariance matrix

$$COV(\boldsymbol{\eta}) = \begin{pmatrix} \mathbf{I}_4 & \mathbf{R} & \mathbf{0} \\ \mathbf{R} & \mathbf{I}_4 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_3 \end{pmatrix}$$

3.4 Results

For complete data, Table 1 lists the results from both approaches. The canonical correlations, the estimates of the weights of X and Y are similar from both approaches due to rounding errors (sometimes with a different sign but still provides the same results). There are 4 canonical correlations as the dimension of the smaller set of variables is 4. Ordered from the largest to the smallest, they are 0.9043192, 0.6146330, 0.6006710, and 0.2170920. In SEM, the canonical correlations are shown as covariance between different pair of latent variables (η) with unit variance. After re-organizing (2, 4, 1, 3 for X and 6, 8, 5, 7 for Y), we have

	Estimate	Std.Err	Z-value	P(> |z|)
eta2 ~~ eta6 :	0.904	0.031	28.940	0.000
eta4 ~~ eta8 :	0.615	0.107	5.760	0.000
eta1 ~~ eta5 :	0.601	0.110	5.480	0.000
eta3 ~~ eta7 :	0.217	0.163	1.328	0.184

In addition to estimates, SEM software also provides their corresponding standard errors, z values and p values. Based on the order of canonical correlations, the re-organized loading matrix of X is obtained as follow.

$$\Lambda_X = \begin{pmatrix} 3.074 & 1.836 & 0.916 & -2.485 \\ -7.134 & 13.043 & 0.462 & 5.244 \\ 3.991 & 2.308 & 2.109 & 4.071 \\ 12.236 & 3.385 & -3.207 & 3.163 \end{pmatrix}.$$

The weight matrix of X is obtained as the transpose of the inverse of loading matrix.

$$W_Y = \begin{pmatrix} 0.05990199 & 0.10610786 & 0.14778630 & -0.195442661 \\ -0.02548197 & 0.06205122 & -0.02765598 & 0.004129405 \\ 0.03901400 & -0.03021496 & 0.22396835 & 0.108494720 \\ 0.03909513 & 0.01937608 & -0.12630358 & 0.016120259 \end{pmatrix}.$$

For Y, the re-organized loading matrix is

$$\Lambda_Y = \begin{pmatrix} 26.061 & 23.883 & -7.307 & -9.992 & 9.991 & 5.611 & 4.853 \\ 19.459 & -1.021 & -7.736 & -6.402 & -4.176 & -5.595 & 6.842 \\ -1.842 & 3.436 & -6.755 & 12.896 & -1.243 & -5.476 & -0.129 \\ -22.282 & 2.228 & 2.931 & -6.224 & -15.675 & 8.323 & 9.280 \\ -8.784 & -4.043 & -32.032 & -9.451 & -33.259 & -10.941 & -16.060 \\ 7.207 & 6.005 & 29.614 & -10.736 & -5.185 & -23.932 & 0 \\ 197.238 & 57.688 & -74.403 & 10.947 & -218.822 & 0 & 0 \end{pmatrix}$$

And its weight matrix is the previous four columns of the following matrix

$$w_Y = \begin{pmatrix} -0.001932080 & 0.03309409 & -0.008146909 & -0.012061048 & 0.009149787 & 0.001069227 & -0.003604512 \\ 0.008238362 & -0.02996765 & -0.022407600 & -0.013311480 & 0.006478400 & -0.028198185 & 0.061358140 \\ -0.014339519 & 0.02883383 & -0.018446419 & 0.042034528 & 0.003051289 & -0.039427278 & 0.033180670 \\ -0.017323157 & 0.01271601 & -0.001538340 & -0.006735220 & -0.012076011 & 0.001708119 & 0.037150307 \\ -0.006968780 & 0.00434944 & -0.012748814 & -0.013545125 & -0.001477576 & -0.010386398 & -0.016015302 \\ -0.001936593 & 0.01060145 & 0.012842720 & -0.005483783 & -0.003591782 & -0.020578041 & -0.002540349 \\ 0.002182018 & 0.00009596 & 0.001904032 & 0.002135723 & -0.003118395 & 0.002754796 & -0.001690846 \end{pmatrix}$$

For incomplete data sets, Tables 2 and 3 provide the comparison between the two approaches. From these tables, we can see the ML estimates are closer to the true estimates of complete data than those from conventional CCA, especially in the case with a low missing data rate. It is because the conventional CCA approach deletes more variables when there are missing values, but SEM approach uses all available values.

4 Conclusions and Discussion

This article investigates the structural equation modeling approach to canonical correlation analysis when data have missing values. By using SEM, missing data can be easily handled with maximum likelihood (ML) estimation method under the MCAR or MAR assumption. It is very practical for researchers to run any SEM software to conduct canonical correlation analysis.

Appendix

See Tables 1, 2 and 3.

Table 1 The canonical correlations (CCs) and the weights of **X** and **Y** obtained from the conventional CCA and from the SEM approach for complete data

Complete data without any missing values				
	CC 1	CC 2	CC 3	CC 4
Conventional CCA approach				
CCs	0.9043192	0.6146330	0.6006710	0.2170920
X weight coefficients				
x1	−0.05902017	0.10451992	−0.14558750	0.19251424
x2	0.02510597	0.06112781	0.02725057	−0.00407127
x3	−0.03844114	−0.02975079	−0.22066721	−0.10688032
x4	−0.03851303	0.01908197	0.12443830	−0.01588451
Y weight coefficients				
y1	0.001903454	0.0326031166	0.008026194	0.011882483
y2	−0.008116932	−0.0295245314	0.022075638	0.013113633
y3	0.014126789	0.0284056630	0.018173654	−0.041410665
y4	0.017065997	0.0125260479	0.001515223	0.006635931
y5	0.006865899	0.0042858813	0.012559922	0.013344329
y6	0.001908078	0.0104446401	−0.012652317	0.005402780
y7	−0.002149688	0.0000945811	−0.001875860	−0.002104078
[1]SEM approach				
CCs	0.904	0.615	0.601	0.217
X weight coefficients				
x1	0.05990199	0.10610786	0.14778630	−0.195442661
x2	−0.02548197	0.06205122	−0.02765598	0.004129405
x3	0.03901400	−0.03021496	0.22396835	0.108494720
x4	0.03909513	0.01937608	−0.12630358	0.016120259
Y weight coefficients				
y1	−0.001932080	0.03309409	−0.008146909	−0.012061048
y2	0.008238362	−0.02996765	−0.022407600	−0.013311480
y3	−0.014339519	0.02883383	−0.018446419	0.042034528
y4	−0.017323157	0.01271601	−0.001538340	−0.006735220
y5	−0.006968780	0.00434944	−0.012748814	−0.013545125
y6	−0.001936593	0.01060145	0.012842720	−0.005483783
y7	0.002182018	0.00009596	0.001904032	0.002135723

Note

1. We use the R package “Lavaan”. The results obtained from other EM software (such as Mplus, EQS, and AMOS) are similar

2. The number of participants is 34, with 11 variables per participant. No missing value

Table 2 The canonical correlations (CCs) and the weights of X and Y obtained from the conventional CCA and from the SEM approach for incomplete data with a low missing values of 2/374 = 0.53%

Incomplete data set with a low missing data rate				
	CC 1	CC 2	CC 3	CC 4
Comparison	0.9043192	0.6146330	0.6006710	0.2170920
Conventional CCA approach (list-wise deleted 2 rows)				
CCs	0.9087407	0.6219618	0.6010693	0.1897089
X weight coefficients				
x1	−0.06327686	0.17569498	0.004242995	−0.188044446
x2	0.02134795	0.01551462	−0.067030627	0.001588004
x3	−0.03586083	0.15391983	0.148485233	0.109242135
x4	−0.03866504	−0.08987870	−0.082728889	0.013705469
Y weight coefficients				
y1	0.001561968	0.009242501	−0.029999995	−0.012767573
y2	−0.010386245	−0.039743075	0.008410371	0.001850001
y3	0.011987609	0.002319815	−0.036737540	0.045260595
y4	0.015148173	0.003549510	−0.014191275	−0.001864996
y5	0.007087851	−0.007386586	−0.010261669	−0.016143496
y6	0.001289599	0.016196617	−0.002156951	−0.006530064
y7	−0.002102193	0.002134729	0.001071663	0.001353122
[1]SEM approach (maximum likelihood estimation)				
CCs	−0.908	−0.610	0.601	−0.221
X weight coefficients				
x1	0.06751507	0.08217289	0.15733362	0.191548052
x2	−0.02402932	0.06672984	−0.01627857	−0.001123253
x3	0.04048745	−0.06213096	0.21458046	−0.111973065
x4	0.03687253	0.03637601	−0.12244213	−0.015435734
Y weight coefficients				
y1	0.001802673	−0.0338834109	−0.004239946	−0.009487170
y2	−0.008230510	0.0250239975	−0.024062211	−0.022063925
y3	0.013328787	−0.0314414361	−0.012483106	0.034006327
y4	0.016692642	−0.0142463283	0.002086912	−0.012493371
y5	0.007160468	−0.0060517848	−0.012211561	−0.012308026
y6	0.001410465	−0.0089522936	0.014622376	−0.006871182
y7	−0.002219914	0.0002349608	0.001901922	0.002054829

The conventional CCA conducts list-wise deletion by removing 22/374 = 5.88% values

Note

1. We use the R package “Lavaan”. The results obtained from other EM software (such as Mplus, EQS, and AMOS) are similar
2. The number of participants is 34, with 11 variables per participant. There are 2 missing values

Table 3 The canonical correlations (CCs) and the weights of X and Y obtained from the conventional CCA and from the SEM approach for incomplete data with a missing values of 15/374 = 4.01%

Incomplete data set with a medium missing data rate				
	CC 1	CC 2	CC 3	CC 4
Comparison	0.9043192	0.6146330	0.6006710	0.2170920
Conventional CCA approach (list-wise deleted 15 rows)				
CCs	0.9149477	0.8187605	0.5383546	0.2265207
X weight coefficients				
x1	−0.04392676	−0.23192330	0.12099261	0.08540134
x2	0.02505721	−0.03346917	−0.04630575	0.04324767
x3	−0.04887578	−0.08238440	0.04787394	−0.29222197
x4	−0.04661994	0.04619791	−0.07644999	0.06891641
Y weight coefficients				
y1	0.001039012	0.0060461986	−0.0069094429	−0.001279544
y2	−0.015564818	0.0348745930	0.0196639634	−0.022117996
y3	0.008573322	0.0108939840	−0.0246776487	−0.043771060
y4	0.011827655	−0.0068842057	−0.0022723941	0.013402196
y5	0.006898159	0.0239773050	0.0056175616	−0.000405460
y6	0.005220909	0.0007958232	0.0191657900	−0.009592668
y7	−0.001759073	−0.0033829867	0.0001747579	0.002154686
[1]SEM approach (maximum likelihood estimation)				
CCs	0.927	−0.680	0.518	−0.260
X weight coefficients				
x1	0.06701171	0.18565004	−0.07916582	−0.258910462
x2	−0.02552725	0.05235413	0.04158887	−0.000710222
x3	0.04204198	0.05693106	−0.23114617	0.066594721
x4	0.03748437	−0.01937753	0.12101047	0.037188424
Y weight coefficients				
y1	−0.002750679	−0.0261097945	0.024482275	−0.001842484
y2	0.012843298	0.0250536632	0.008895954	0.045124906
y3	−0.011470019	−0.0261439348	−0.001397013	−0.021463728
y4	−0.017037333	−0.0208146911	0.007571625	0.022986906
y5	−0.007151805	0.0044416329	0.018233471	−0.006383037
y6	−0.003264085	−0.0120543186	−0.006458153	0.002366780
y7	0.001938246	−0.0009951285	−0.002568367	−0.002137180

The conventional CCA conducts list-wise deletion by removing 165/374 = 44.12% values

Note

1. We use the R package "Lavaan". The results obtained from other EM software (such as Mplus, EQS, and AMOS) are similar

2. The number of participants is 34, with 11 variables per participant. There are 15 missing values

References

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.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8*(3), 430–457.

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.

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.

Lu, Z. L., & Gu, F. (2018). A structural equation modeling approach to canonical correlation analysis. *Quantitative psychology* (pp. 261–273). Cham: Springer.

MathWorks, Inc. (2012). *MATLAB and statistics toolbox*. Massachusetts: Natick.

Muthén, B. O., & Muthén, L. K. (2012). *Software Mplus version 7*.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. 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 Corporation, Armonk, NY.

Yuan, K.-H., & Lu, L. (2008). SEM with missing data and unknown population distributions using two-stage ML: Theory and its application. *Multivariate Behavioral Research, 43*(4), 621–652.

Small-Variance Priors Can Prevent Detecting Important Misspecifications in Bayesian Confirmatory Factor Analysis

Terrence D. Jorgensen, Mauricio Garnier-Villarreal, Sunthud Pornprasermanit and Jaehoon Lee

Abstract We simulated Bayesian CFA models to investigate the power of PPP to detect model misspecification by manipulating sample size, strongly and weakly informative priors for nontarget parameters, degree of misspecification, and whether data were generated and analyzed as normal or ordinal. Rejection rates indicate that PPP lacks power to reject an inappropriate model unless priors are unrealistically restrictive (essentially equivalent to fixing nontarget parameters to zero) and both sample size and misspecification are quite large. We suggest researchers evaluate global fit without priors for nontarget parameters, then search for neglected parameters if PPP indicates poor fit.

Keywords Structural equation modeling · Confirmatory factor analysis · Bayesian inference · Model evaluation · Model modification

Bayesian structural equation modeling (BSEM) has recently received substantial attention within psychology and the social sciences as an increasingly viable alternative to traditional frequentist SEM techniques (MacCallum, Edwards, & Cai, 2012; Muthén & Asparouhov, 2012; Rindskopf, 2012), such as maximum likelihood (ML) estimation. Bayesian estimates of model parameters are based on a sampling plausible parameter values from the posterior distribution of the model parameters, which is estimated using Markov chain Monte Carlo (MCMC) estimation (see Muthén & Asparouhov, 2012, for details). Programs available for analyzing a BSEM include Amos (Arbuckle, 2012), M*plus* (Muthén & Muthén, 2012), and more recently the R (R Core Team, 2018) package blavaan (Merkle & Rosseel, 2018), which utilizes the

T. D. Jorgensen (✉)
University of Amsterdam, Nieuwe Achtergracht 127, 1018WS Amsterdam, The Netherlands
e-mail: T.D.Jorgensen@uva.nl

M. Garnier-Villarreal
Marquette University, 2340 N. Cramer St., Unit 515, Milwaukee, WI 53211, USA
e-mail: mauricio.garnier@marquette.edu

S. Pornprasermanit · J. Lee
Texas Tech University, 41071, Lubbock, TX 79409, USA
e-mail: jaehoon.lee@ttu.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_23

more general Bayesian modeling software JAGS (Plummer, 2003) and Stan (Carpenter et al., 2017).

An important step in fitting BSEMs (or models in general) is to investigate how well the hypothesized model can reproduce the observed data, for which posterior predictive model checking (PPMC; Gelman, Meng, & Stern, 1996) was developed (see Levy, 2011, for a review of PPMC specifically for BSEM). PPMC is not a test statistic per se, but it can be based on a test statistic, such as the traditional χ^2 test of exact fit derived from the ML discrepancy function (although any discrepancy measure of interest can be used, such as the SRMR fit index; Levy, 2011). PPMC uses the samples from the joint posterior distribution to check data–model correspondence by comparing observed data to expected data. For the sampled parameters at one iteration of a Markov chain (after the burn-in iterations), a random sample of N observations is simulated from the population with those parameters. The simulated sample's sufficient statistics (means and covariances) are calculated and compared (using the ML discrepancy function) with the sufficient statistics implied by the model parameters drawn from the posterior distribution at that iteration; likewise, the same discrepancy is calculated comparing the sufficient statistics of the observed data to the model-implied sufficient statistics at that iteration. This results in a discrepancy measure both for the observed data and for the simulated data. A score of 1 is assigned if the observed data have less discrepancy (i.e., fit better) than the simulated data; otherwise, a score of 0 is assigned—this score can be considered as a Bernoulli random variable. These binary numbers are averaged across all iterations sampled from the posterior distribution. The average is referred to as the posterior predictive *p* value (PPP).

Naturally, the model should fit well to the simulated data at every iteration because simulated data are drawn from those population parameters. If the hypothesized model is an appropriate approximation of the population process from which the real data was sampled, then the model should fit the real data often, too. On average, an appropriate model will fit the real data better than the simulated data about as often as the other way around, so the mean of this Bernoulli random variable is 50%, the expected value of PPP when the target model is approximately correct. The probability will decrease as the appropriateness of the hypothesized model decreases in its ability to explain the phenomena under investigation. That is, if the model is grossly inappropriate, the model will continue to fit well to the simulated data drawn from those population parameters, but it will rarely fit well to the real data, so the expected value of PPP will approach 0%.

There is no theoretical cutoff for how low PPP must be to indicate unignorable misfit, nor is there a consensus about how applied users should interpret PPP (e.g., treat it like a frequentist *p* value and compare it to an alpha level, or use as a fit index). Muthén and Asparouhov (2012) recently suggested after their initial simulations that the traditional approach of using "posterior predictive *p* values of 0.10, 0.05, or 0.01 appears reasonable" (p. 315), with the caveat that further investigations were needed to establish how these methods behave in practice with various models and data. Depaoli (2012) already began to address that gap in a recent Monte Carlo simulation study of PPP values in the context of growth mixture modeling, and found that PPP and graphical PPMC were only likely to identify extremely misspecified growth mixture models. We use a Monte Carlo simulation study to investigate (a) the sensitivity of PPP to varying levels of misspecification in confirmatory factor analysis (CFA) models, as well as (b) how consistently Muthén and Asparouhov's guidelines would apply across varying samples sizes and different informative priors.

1 Method

1.1 Continuous (Standard Normal) Indicators

Using the MONTECARLO command in M*plus* (version 6.11 for Linux; Muthén & Muthén, 2002), we simulated a two-factor CFA with three indicators per factor. In each of the four population models, factors were standard normal ($\mu = 0, \sigma = 1$), with a factor correlation = 0.25, factor loadings = 0.70, indicator intercepts = 0.0, and indicator residual variances = 0.51; thus, indicators had unit variance. To vary levels of misspecification of the analysis model, the third indicator of the first factor was specified to have a cross-loading on the second factor (λ_{32}) in the population. The magnitude of λ_{32} was 0.0, 0.2, 0.5, or 0.7 in the population, but was constrained to be close to zero in the analysis model using informative priors. For ease of interpretation, we refer to $\lambda_{32} = 0.2$ as minor misspecification (using $\alpha = 0.05$, the ML χ^2 test has 80% power when $N > 500$, RMSEA = 0.06, SRMR = 0.03, CFI = 0.98), $\lambda_{32} = 0.5$ as severe misspecification (80% power when $N > 150$, RMSEA = 0.12, SRMR = 0.07, CFI = 0.92), and $\lambda_{32} = 0.7$ as very severe misspecification (80% power when $N > 100$, RMSEA = 0.14, SRMR = 0.07, CFI = 0.89).

In the analysis model, we specified noninformative priors for all target parameters (primary loadings, residual variances, and the factor covariance) using M*plus* defaults—for example, factor loadings ~N($\mu = 0$, $\sigma^2 =$ "infinity"). For all cross-loadings, we specified normally distributed priors with four levels of informative variance, chosen to correspond approximately with the prior belief in a 95% probability that the cross-loadings are within approximately ±0.01, ±0.10, ±0.20, or ±0.30 of zero (i.e., $\sigma = 0.005$, 0.05, 0.10, and 0.15, or equivalently $\sigma^2 = 0.000025$, 0.0025, 0.01, and 0.0225). In each condition, sample sizes of $N = 50$–500 were drawn in increments of 25, along with an asymptotic condition of $N = 1000$. We drew 200 samples from each of 320 conditions (20 sample sizes, four levels of CL, and four prior variances) with normally distributed indicators.

Following Muthén and Asparouhov's (2012) advice, we kept 100,000 iterations from the MCMC chains after thinning every 100th iteration. Over 99% of models converged on a proper solution, yielding 63,480 (out of 64,000) PPP values for analysis. Convergence in each condition was at least 98% except when sample size was small ($N \leq 100$) and CL was large ($\lambda_{32} \geq 0.5$). The smallest convergence rate was 82% ($N = 50$, $\lambda_{32} = 0.7$).

1.2 Categorical Indicators

Because behavioral data are so often measured using discrete scales rather than truly continuous data, we also simulated binary and ordinal data. Rhemtulla, Brosseau-Liard, and Savalei's (2012) simulation results suggest that when ordinal variables have at least five categories, robust estimation methods for nonnormal data provide similar conclusions as estimation for categorical data. For few categories, we were interested in how PPP and constrained nontarget parameters would be affected by the data distribution. Thus, we manipulated the same conditions as for normal indicators described above, but we additionally varied the number of categories (from two to five) and how the data were analyzed (appropriately as ordinal or inappropriately as normal). Thresholds were not manipulated, but were chosen to mimic a unimodal symmetric distribution: 0 for binary; ± 0.8 for three categories; −1, 0, and 1 for four categories; and −1.6, −0.8, 0.8, and 1.6 for five categories. Whereas 100% of the models converged when the indicators were analyzed as ordinal, when the indicators were analyzed as normal, convergence rates were 94.78, 97.34, 98.45, and 98.13% for 2, 3, 4, and 5 categories, respectively.

2 Results

For each model, we investigate the sampling variability of PPP across conditions, and calculate power and Type I error[1] rates using traditional cutoff values ($\alpha = 0.10$, 0.05, or 0.01) for PPP to identify "significant" misfit.

2.1 Sampling Variability of PPP

Table 1 shows the effect sizes for each model under investigation. Using Cohen's (1988) criteria for interpreting the size of η^2 (negligible < 0.01 < small < 0.06 < moderate < 0.14 < large), N had a negligible effect on PPP when normal or ordinal data were analyzed assuming normality, but N explained 4% of variance in PPP when data were analyzed as ordinal. PPP values were largely influenced by the magnitude of the neglected cross-loading (CL), but much more so for normal data ($\eta^2 = 34.2\%$) than for categorical data analyzed as normal ($\eta^2 = 17.1\%$) or as ordinal ($\eta^2 = 18\%$). The magnitude of prior variance for estimating nontarget CLs had a large effect on PPP when normal ($\eta^2 = 22.1\%$) or categorical ($\eta^2 = 20.9\%$) data were analyzed as normal, but only a moderate effect when categorical data were analyzed as ordinal ($\eta^2 = 7\%$). When categorical data were analyzed as ordinal, the number of cate-

[1] We use the term "Type I error" when referring to any model that does not omit a substantial parameter, although in the categorical data conditions, the model contains another type of misspecification (incorrect likelihood) when analyzed as though the data were normally distributed.

Table 1 Proportions of variance (η^2) of PPP explained by Monte Carlo factors

	How data were generated (and analyzed)		
	Normal data (as normal)	Categorical data (as normal)	Categorical data (as ordinal)
N	0.002	0.001	0.040
Prior variance	**0.221**	**0.209**	**0.070**
Misfit	**0.343**	**0.171**	**0.180**
Number of categories (#CAT)		0.019	**0.072**
N × Prior	0.014	0.024	0.018
N × Misfit	0.002	0.003	0.021
Prior × Misfit	**0.103**	**0.096**	0.023
N × #CAT		0.000	0.001
Prior × #CAT		0.010	0.008
Misfit × #CAT		0.013	0.017
N × Prior × Misfit	0.009	0.010	0.012
N × Prior × #CAT		0.001	0.001
N × Misfit × #CAT		0.000	0.001
Prior × Misfit × #CAT		0.008	0.003
N × Prior × Misfit × #CAT		0.002	0.001

Note Medium and large effect sizes using Cohen's (1988) criteria (i.e., effect explains at least 6% of variance) are in bold font. Data generated as continuous did not have varying numbers of categories, so cells involving the #CAT effect are blank

gories also had a moderate effect of PPP ($\eta^2 = 7.2\%$). The only substantial two-way interactions were found between prior variance and magnitude of neglected CL, for normal data ($\eta^2 = 10.3\%$) and for categorical data analyzed as normal ($\eta^2 = 9.6\%$). All other interactions effects were negligible or small ($\eta^2 \leq 4\%$). The effect of *N* in most conditions appears approximately linear in Fig. 1 (and Figs. A1–A8 provided in the online[2] supplemental materials), so we treated *N* as a continuous[3] covariate to calculate η^2.

Figure 1 (online supplemental material) reveals the nature of the interaction between magnitude of prior variances and of the neglected parameter (λ_{32}) in normal-data conditions. When $\lambda_{32} = 0$ (no misspecification), the average PPP value is consistent with its expected value of 50%, regardless of the magnitude of prior variance. As the magnitude of the neglected population parameter increases, the average PPP decreases, but PPP shows more sensitivity to misspecification when prior variances are restrictive than when only weakly informative. Figure 1 plots PPP values only

[2] The online supplemental materials can be retrieved at the following URL: https://osf.io/buhvg/.

[3] Treating *N* as a categorical factor showed no substantial difference in the effect sizes.

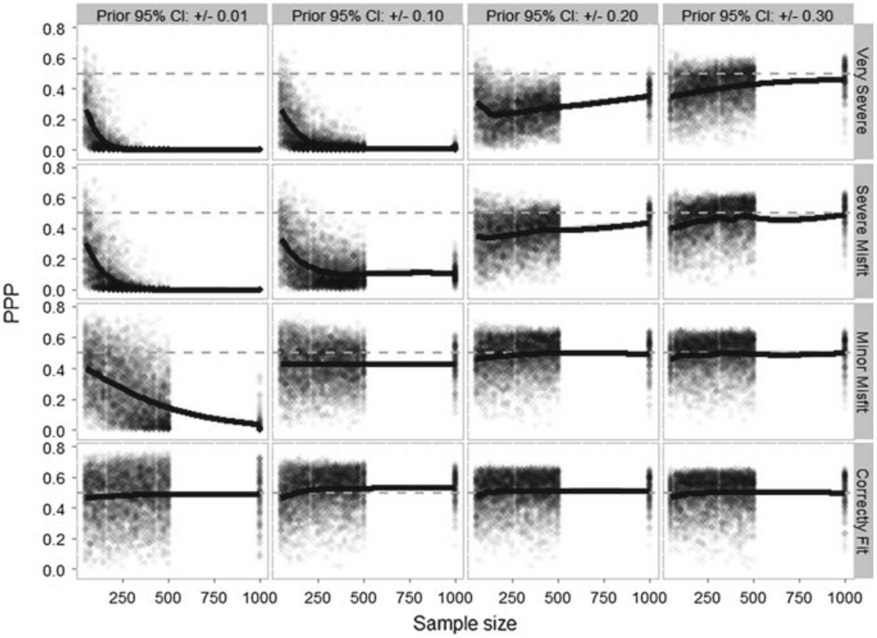


Fig. 1 Variability of PPP as a function of sample size, plotted separately for each condition of prior variance and magnitude of neglected cross-loading (λ_{32}). There is a smoothed regression line indicating how the mean PPP changes, and the horizontal dashed line in each cell at PPP = 50% refers to the expected value under the null hypothesis of no misfit

for the normal-data conditions, but this two-way interaction is characterized by similar patterns in all categorical-data conditions, with the exception that PPP appears less sensitive for two-category data, especially when analyze as ordinal (see Figures A1–A4 in the online appendix). PPP is also less variable when categorical data have fewer categories and are analyzed as ordinal, so PPP's distribution resembles even less the expected uniform [0, 1] distribution of traditional *p* values (Hoijtink & van de Schoot, 2018).

2.2 Detecting Misfit

Figure 2 (and online supplemental material) plots the rejection rates for normal-data models against *N*, with separate panels for each magnitude of prior variance and misspecification. We used three different criteria for rejecting a model due to lack of fit: PPP < 0.10, 0.05, or 0.01, to evaluate Muthén and Asparouhov's (2012, p. 315) suggestion. Rejection rates in the bottom row of Fig. 2 represent Type I error rates because the model is not misspecified (i.e., $\lambda_{32} = 0$), whereas rejection rates in the top two rows represent power (i.e., when $\lambda_{32} = 0.5$ or 0.7). Rejection rates in the

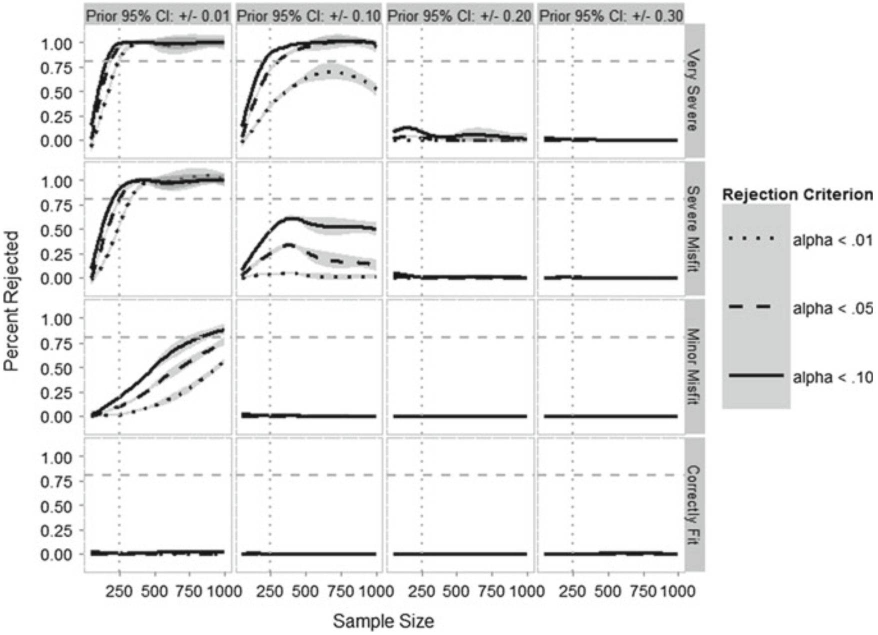


Fig. 2 Rejection rates as a function of sample size, plotted separately across conditions of varying priors and magnitude of neglected cross-loading (λ_{32}). The dashed horizontal line at 80% represents acceptable power, and the dotted vertical line at $N = 250$ is for convenience when judging sample sizes required for adequate power

row labeled "Minor Misfit" correspond to $\lambda_{32} = 0.2$ and could be classified as Type I errors or power, depending on the whether the analyst wishes to test exact fit or close fit (Browne & Cudeck, 1992). That is, a researcher might consider a neglected cross-loading of 0.2 to be of little substantive consequence, so the analysis model would be considered to correspond closely enough to the population model that it should not be rejected.

Consistent with prior research showing that PPP is more conservative than nominal error rates (Gelman et al., 1996; Levy, 2011), the Type I error rate is near zero in almost every condition, much lower than nominal levels using any of the three rejection criteria. However, power rarely exceeds 80% (a commonly preferred minimum) unless the neglected parameter is quite large or the prior variance is quite small (or both). Depending on the rejection criterion, power exceeds 80% when $N > 200$–300 when using the most restrictive priors. This suggests that with sufficient sample size, researchers could only be confident about detecting misfit by specifying priors so informative that their 95% confidence limits are approximately ±0.01—so strongly informative that the model bears little practical distinction from one in which no informative priors are specified for cross-loadings. Using more realistic informative priors with 95% confidence limits approximately ±0.1, power only exceeded 80% for the most severe level of misspecification ($\lambda_{32} = 0.7$). Perhaps most noteworthy, power was close to zero to detect severe misspecification ($\lambda_{32} = 0.5$) at any N when using Muthén and Asparouhov's (2012) suggested priors (95% confidence limits approximately ±0.2). Similar results were found for categorical data (see Figs. A9–A16), although power was even lower when data with fewer categories were analyzed as ordinal (e.g., power remained nearly zero in all binary conditions; see Fig. A9).

3 Discussion

The assessment of fit and detection of misspecification in SEM is no less important in a Bayesian context than in a traditional frequentist paradigm, but tools currently available in BSEM are few, and their behavior is largely unknown. In the conditions we investigated, PPP lacks power unless $N > 200$–300, misspecification is severe, and priors for nontarget parameters are highly (even unrealistically) restrictive. This implies that informative priors for nontarget parameters should be chosen very carefully. Asparouhov, Muthén, and Morin (2015) suggested a data-driven sensitivity analysis to choose priors that balanced detecting substantial misfit and allowing negligible misfit. More recently, Cain and Zhang (in press) found larger Type I error rates with simulated 3-factor models than we did with 2-factor models, and they recommended different PPP criteria for models with different numbers of indicators. This calls into question whether any uniform cutoffs can be expected to perform consistently across conditions with different data and model characteristics. Because power to detect substantial misspecification only appears adequate when priors are so restrictive that they are nearly equivalent to fixing the nontarget parameters to zero, we suggest researchers simply evaluate global fit without priors for nontarget parameters, then search for neglected parameters only if PPP indicates poor fit. But because even minor misspecification can be detected with great power in asymptotically large samples (Hoofs, van de Schoot, Jansen, & Kant, 2018), the development of complementary fit indices for evaluating BSEMs (similar to those used in SEM) would be a welcome and useful addition to the researcher's toolbox. The BRMSEA (Hoofs et al., 2018) is the first such attempt, and it appears promising, but further development is needed.

References

Arbuckle, J. L. (2012). *IBM SPSS Amos 21 user's guide*. Chicago, IL: IBM.

Asparouhov, T., Muthén, B., & Morin, A. J. S. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeyer et al. *Journal of Management, 41*, 1561–1577. https://doi.org/10.1177/0149206315591075.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research, 21*(2), 230–258. https://doi.org/10.1177/0049124192021002005.

Cain, M. K., & Zhang, Z. (in press). Fit for a Bayesian: An evaluation of PPP and DIC for structural equation modeling. *Structural Equation Modeling*. https://doi.org/10.1080/10705511.2018.1490648.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Depaoli, S. (2012). The ability for posterior predictive checking to identify model misspecification in Bayesian growth mixture modeling. *Structural Equation Modeling, 19*(4), 534–560. https://doi.org/10.1080/10705511.2012.713251.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*, 733–807. https://doi.org/10.1.1.142.9951.

Hoijtink, H., & van de Schoot, R. (2018). Testing small variance priors using prior-posterior predictive *p* values. *Psychological Methods, 23*(3), 561–569. https://doi.org/10.1037/met0000131.

Hoofs, H., van de Schoot, R., Jansen, N. W., & Kant, I. (2018). Evaluating model fit in Bayesian confirmatory factor analysis with large samples: Simulation study introducing the BRMSEA. *Educational and Psychological Measurement, 78*(4), 537–568. https://doi.org/10.1177/0013164417709314.

Levy, R. (2011). Bayesian data–model fit assessment for structural equation modeling. *Structural Equation Modeling, 18*(4), 663–685. https://doi.org/10.1080/10705511.2011.607723.

MacCallum, R. C., Edwards, M. C., & Cai, L. (2012). Hopes and cautions in implementing Bayesian structural equation modeling. *Psychological Methods, 17*(3), 340–345. https://doi.org/10.1037/a0027131.

Merkle, E. C., & Rosseel, Y. (2018). Blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software, 85*(4), 1–30. https://doi.org/10.18637/jss.v085.i04.

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313–335. https://doi.org/10.1037/a0026802.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Retrieved from http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/.

R Core Team. (2018). R: A language and environment for statistical computing (version 3.5.1) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from the comprehensive R archive network (CRAN): https://www.R-project.org/.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373. https://doi.org/10.1037/a0029315.

Rindskopf, D. (2012). Next steps in Bayesian structural equation models: Comments on, variations of, and extensions to Muthen and Asparouhov (2012). *Psychological Methods, 17*(3), 336–339. https://doi.org/10.1037/a0027130.

Measuring the Heterogeneity of Treatment Effects with Multilevel Observational Data

Youmi Suk and Jee-Seon Kim

Abstract Multilevel latent class analysis and mixture propensity score models have been implemented to account for heterogeneous selection mechanisms and for proper causal inference with observational multilevel data (Kim & Steiner in Quantitative Psychology Research. Springer, Cham, pp. 293–306, 2015). The scenarios imply the existence of multiple selection classes, and if class membership is unknown, homogeneous classes can be usually identified via multilevel logistic latent class models. Although latent class random-effects logistic models are frequently used, linear models and fixed-effects models can be alternatives for identifying multiple selection classes and estimating class-specific treatment effects (Kim & Suk in Specifying Multilevel Mixture Models in Propensity Score Analysis. International Meeting of Psychometric Society, New York, 2018). Using the Korea TIMSS 2015 eighth-grade student data, this study examined the potentially heterogeneous treatment effects of private science lessons by inspecting multiple selection classes (e.g., different motivations to receive the lessons) using four types of selection models: random-effects logistic, random-effects linear, fixed-effects logistic, and fixed-effects linear models. Implications of identifying selection classes in casual inference with multilevel assessment data are discussed.

Keywords Causal inference · Multilevel propensity score matching · Finite mixture modeling · Latent class analysis · Selection bias · Balancing scores · Heterogeneous selection processes · Heterogeneous treatment effects · Hierarchical linear modeling

Y. Suk (✉) · J.-S. Kim
Department of Educational Psychology, Educational Sciences Building, University of Wisconsin-Madison, 1025 West Johnson Street, Madison, WI 53706, USA
e-mail: ysuk@wisc.edu

J.-S. Kim
e-mail: jeeseonkim@wisc.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_24

1 Heterogeneous Treatment Effects in Assessment Data

Measuring causal treatment effects has been studied extensively both in randomized control trials and observational studies. The term “treatment” came from early experiments in the fields of agriculture and medicine, and is now widely used in different fields of natural and social sciences, including education, psychology, economics, and political science. A treatment (e.g., therapy, private tutoring) is provided to some units, called a treated/treatment group, but not others, called an untreated/control/comparison group (Holland, 1986; Imai, King, & Stuart, 2008). In the randomized experimental design, treatment assignment is controlled by researchers. This means that there is no treatment selection bias in the design if treatment assignment is implemented in a proper way. However, in observational studies, treatment selection is determined by factors beyond researchers' control—*why did some students receive the treatment and not others? Does the treated group have any systematic difference from the untreated group?* This selection bias is a main obstacle to estimating treatment effects (e.g., difference in outcome between the treated group and untreated group). To remove the selection bias, we can use propensity score (PS) analysis; PS is one of the balancing scores that allow the conditional distribution of observed covariates to be identical for treated and untreated groups (Rosenbaum & Rubin, 1983). PS is defined as a unit's conditional probability of belonging to the treatment group given observed covariates, and it is estimated most commonly in logistic regression models.

Discovering heterogeneous treatment effects has been necessary when subpopulations have differential gains after receiving a treatment in observational assessment data. Treatment effects may differ from individual to individual and from subgroup to subgroup. In case we suspect heterogeneous treatment effects, a small sample size may hinder the estimation of treatment effect heterogeneity (Wager & Athey, 2018). However, large-scale assessment data, such as the National Assessment of Educational Progress (NAEP), the Program for International Student Assessment (PISA), and the Trends in International Mathematics and Science Study (TIMSS), tends to come with large sample sizes, so sample size may not be an obstacle to exploring treatment effect heterogeneity. Another obstacle to the estimation of heterogeneous treatment effects in assessment data comes from the heterogeneity of treatment selection bias. Since treatment selection in observational assessment data is not random and could be heterogeneous, we need to eliminate selection bias that could vary depending on subpopulations (Kim & Steiner, 2015). For example, students in some schools receive the treatment of private lessons because they easily have access to the lessons, whereas students in other schools receive them mainly due to the poor quality of school lessons. These distinct selection scenarios imply the existence of multiple selection groups and produce heterogeneous treatment effects. Furthermore, when assessment data has clustered structures, it requires additional corrections to accurately remove the selection bias. In a way, mixture multilevel PS models allow us to account for both heterogeneity and intra-cluster correlation.

To capture the heterogeneity of average treatment effects (ATE) using multilevel assessment data, multilevel PS matching techniques have gained popularity (e.g., Kim & Suk, 2018; Leite, 2016). When we suspect heterogeneous selection processes, we can apply within-class matching (Kim & Steiner, 2015; Kim, Steiner, & Lim, 2016), which requires matches (between the treated group and untreated group) across clusters but within homogeneous groups of clusters, referred to as "classes." When the class membership is unknown (e.g., unknown selection groups of private lessons), unknown classes can be identified via finite mixture modeling (Clogg, 1995; McLachlan & Peel, 2000) where a latent class random-effects logistic regression model is commonly used (Kim & Steiner, 2015).

A few studies explored alternative models besides latent class random-effects logistic regression models, in order to improve the quality of classification and estimate unbiased, heterogeneous treatment effects (e.g., Kim & Suk, 2018; Suk & Kim, 2018). When using within-class matching, the choice of Level-1 model (e.g., logistic vs. linear) and Level-2 model (e.g., random effects vs. fixed effects) may influence the estimation of heterogeneous treatment effects through a series of analysis procedures (e.g., extraction of the number of latent classes, classification, ATE estimation). Therefore, to demonstrate, this study investigated heterogeneous treatment effects, particularly the effects of private science lessons, with the Korea TIMSS 2015 data by using different multilevel mixture selection models, where a latent class random-effects logistic model is the baseline.

2 Multilevel Mixture Selection Models in Propensity Score Analysis

Multilevel PS matching techniques can be explained using a multilevel matching continuum. The two extremes of the continuum are within-cluster matching and across-cluster matching. Within-cluster matching requires matches between treated and untreated units for each cluster separately, while across-cluster matching requires matches for all clusters at once (Hong & Raudenbush, 2006). With sufficient overlap of PS within clusters, within-cluster matching would be suitable, while without sufficient overlap, across-cluster matching can be used. However, pooling all the clusters may be harmful if distinctive selection processes exist across clusters.

As a continuum between the two extremes, Kim and Steiner (2015) proposed within-class matching, incorporating the advantages of within-cluster matching and across-cluster matching and remedying their shortcomings. The basic idea is to identify homogeneous groups of clusters regarding the selection model and to pool units across clusters but within homogeneous groups of clusters or "classes." If membership is known, we use manifest classes (e.g., types, regions), and if membership is unknown, we can identify homogeneous latent classes using mixture modeling (Kim et al., 2016).

Within-class matching strategy gains three obvious advantages against within-cluster and across-cluster matching. First, the selection processes become more homogeneous within classes than across all clusters; therefore, it is better to capture the heterogeneity of the true selection models by using the class-specific PS model specifications. Second, more overlap of PS can be achieved within homogeneous classes than within clusters because within-class matching can find matches across clusters that belong to the same class. Third, we can directly examine heterogeneous treatment effects depending on heterogeneous selection classes because distinctive selection mechanisms may lead to treatment effect heterogeneity (Kim et al., 2016).

It is true that latent classes can be identified at either the cluster level or individual level. When we classify clusters into classes, called cluster-level classification, strictly nested structures are formed in data: individuals nested within clusters, nested within classes. We can also classify individuals into classes, called individual-level classification. The model specification of individual-level classification changes as clusters are unlikely to be nested within classes and individuals of each cluster can belong to different classes. The multilevel mixture approach for cluster-level and individual-level classifications requires different assumptions and modeling. This paper centers on cluster-level classification where clusters are strictly nested within classes (Kim et al., 2016).

When we do cluster-level classification using a latent class random-effects logistic model as a selection model, the model can be written as follows:

$$\operatorname{logit}\left(\pi_{ijs}\right) = \alpha_s + X'_{ijs}\beta_{js} + W'_{js}\gamma_s + XW'_{ijs}\delta_{js} + T_{js} \tag{1}$$

where $i = 1, \ldots, n_{js}$, $j = 1, \ldots, M_s$, $s = 1, 2$ denote individual, cluster, and class, respectively; π_{ijs} is the propensity of belonging to the treatment group for an individual i in cluster j in class s; X_{ijs}, W_{js}, and XW_{ijs} are individual-level covariates, cluster-level covariates, and their cross-level interactions, respectively; T_{js} are random effects for cluster j; α_s are class-specific intercepts; β_{js}, γ_s and δ_{js} are class-specific individual-level regression coefficients, cluster-level regression coefficients, and interaction coefficients, respectively.

Recently, alternatives to random-effects logistic models have been explored in multilevel mixture selection models (Suk & Kim, 2018; Kim & Suk, 2018). These two studies provided evidence that, through simulation studies, linear regression models determined the number of latent classes more accurately than logistic models did. Extracting latent classes correctly is very important when using the within-class matching strategy, because treatment effects based on incorrectly extracted subpopulations are unlikely to be informative. Kim and Suk (2018) also revealed that fixed-effects models outperformed random-effects logistic models regarding the unbiased estimation of ATE with a small cluster size of 20 on average. These results indicate that although random-effects logistic models have been used as a common extension of logistic regression with multilevel-structured data, alternative models produce different benefits depending on the specific conditions of given data. For real data study with unknown selection mechanisms, specifying different models and comparing results are of help to assess the validity of assumptions and reinforce our inference.

3 TIMSS Data Analysis: The Effects of Private Science Lessons

This current study investigated the heterogeneity of the ATE of private science lessons on students' science achievement scores with TIMSS data. To demonstrate, we compared three alternatives—random-effects linear, fixed-effects logistic, fixed-effects linear—with random-effects logistic models for multilevel mixture selection models.

3.1 Korea TIMSS 2015 Data

TIMSS, sponsored by the International Association for the Evaluation of Educational Achievement (IEA), is an educational international study investigating students' achievement progresses in mathematics and science. TIMSS has been conducted for students at Grades 4 and 8, every four years, since 1995 first administered the test to more than 40 countries. The most recent TIMSS data were collected in 2015 across more than 60 countries and other education systems. The data are from a two-stage stratified cluster sample. Specifically, schools were selected first given important demographic variables (e.g., in Korea, their location and/or their gender type), and then one or more intact classrooms were randomly selected within each school (Martin, Mullis, & Hooper, 2016).

Using Korea TIMSS 2015 data of 8th graders, we examined the heterogeneous effects of private science lessons for selection classes across schools since we suspected different selection processes. The data included 5309 students from 150 middle schools where the school sizes varied and the minimum number of students amounted to only six (a range of 6–75; mean 35.4; median 32). We deleted cases of students who gave inconsistent responses regarding their attendance of private science lessons and who had missing values in 12 covariates used in selection models: six student-level covariates and six school-level covariates. Student-level covariates included the sex of students (*Male*), their fathers' highest education level (*Dad.edu*, with three levels of no college graduates, college graduates, and no idea), the number of books at home (*Books25*, with two levels of greater than 25, and less than or equal to 25), the number of home study supports (*Home.spprt*, with three levels of 0, 1, and 2), students' confidence in science (*Stu.conf.sci*), and value in science (*Value.sci*). School-level covariates included school type by gender (*Gender.type*, with three levels of all-boys, all-girls, co-education), the percentage of economically disadvantaged students (*Pct.disad*, with four levels of 0–10%, 11–25%, 26–50% and more than 50%), the area of school location (*City.size*, with four levels of urban, suburban, medium size city, and small town), school emphasis on academic success (*Aca.emph*), instruction affected by science resource shortage (*Res.short*), and school discipline problems (*Dscpn*). As a result, the final sample was 4875 students (91.8% of the initial) from 149 schools.

3.2 Methods

With the average school size being 35, it is hard to achieve sufficient overlap within clusters for capturing the distinctive effects within schools. Therefore, we applied within-class matching to estimate the heterogeneous effects depending on homogeneous selection classes via four types of multilevel latent class modeling: random-effects logistic (RE LOGIT), random-effects linear (RE LINEAR), fixed-effects logistic (FE LOGIT), and fixed-effects linear models (FE LINEAR).

We compared four estimation models regarding the extraction of latent classes, classification, and the performance of ATE estimates. Among a variety of PS techniques (e.g., matching, stratification, inverse-propensity weighting), we applied the marginal mean weighting through stratification (MMW-S), suggested by Hong and Hong (2009). Because some of the predicted values from linear probability models were outside the range of (0, 1), inverse-propensity weighting could not be applied in those cases. Instead, stratification techniques can be used due to the robustness to nonsense prediction. Therefore, we used MMW-S as our balancing score adjustment in this study.

The model specification of RE LOGIT is the same as in the data generating model in Eq. (1), and RE LINEAR's specification is as follows:

$$\text{RE LINEAR:}\quad Z_{ijs} = \alpha_s + X'_{ijs}\beta_{js} + W'_{js}\gamma_{js} + T_{js} + \epsilon_{ijs} \tag{2}$$

where ϵ_{ijs} are random errors for an individual i in cluster j in class s, and other notations are the same as in Eq. (1).

FE LOGIT and FE LINEAR specifications are:

$$\text{FE LOGIT:}\quad \text{logit}\left(\pi_{ijs}\right) = \alpha_s + X'_{ijs}\beta_{js} + D'_{js}\gamma_s + T_{js} \tag{3}$$

$$\text{FE LINEAR:}\quad Z_{ijs} = \alpha_s + X'_{ijs}\beta_{js} + D'_{js}\gamma_s + T_{js} + \epsilon_{ijs} \tag{4}$$

where D_{js} is a dummy variable for each cluster $j > 1$ (omitting the last cluster because of singularity), and there are no level-2 covariates, W_{js}. Other notations are the same with Eqs. (1) and (2).

To extract the optimal number of latent classes, we compared the model fit indices (here, Akaike information criterion; AIC) by increasing the number of latent classes in each multilevel mixture selection model. We used *Mplus8* (Muthén & Muthén, 1998–2017) and R with the *flexmix* package to identify latent classes, and with the *lme4* package to estimate class-specific balancing scores via four selection models. To check covariate balance, as a rule of thumb, we considered absolute standardized mean differences (between the treated and control groups) smaller than 0.1 standard deviations (SD) and variance ratios greater than 4/5 and smaller than 5/4 as good balance for each covariate (Steiner & Cook, 2013).

We used the first plausible value of science achievement scores as our outcome variable, as the purpose of this paper was to apply within-class matching with different multilevel mixture selection models in practice rather than to interpret the estimates of parameters. Much more care would be required regarding plausible values, final sample weights, and replicate weights to interpret descriptive statistics and parameter estimates (Martin et al., 2016).

3.3 Applications of Multilevel Mixture Selection Models

After checking the optimal number of latent classes using four different models, one homogeneous model was preferred over multiple-class models based on FE LOGIT and FE LINEAR models, while two-class models were preferred based on RE LOGIT and RE LINEAR. Because FE models underestimated the number of latent classes with small cluster sizes according to the simulation study of Kim and Suk (2018), we chose two latent classes, relying more on the results of the two RE models.

Assuming the existence of two selection classes, we investigated class membership across different selection models. Class sizes differed from model to model as seen from Table 1, and consistency rates varied across models. Since we do not know true class membership, it was hard to solve the label switching problem, but we determined class labels of each model in a way that optimized the global consistency across models. As a result, we found that RE LINEAR and FE LINEAR resulted in similar class sizes, whereas the class sizes for RE LOGIT and FE LOGIT were unevenly distributed with the first class being rather small in RE LOGIT and the reverse being true for FE LOGIT. In addition, the class membership of FE LOGIT was least consistent with other models' class membership. Besides FE LOGIT, the other three models showed around 70% consistency rates.

Table 2 summarizes the descriptive statistics of selection classes identified by the four model approaches. Means are given for continuous variables, while proportions are provided for binary or categorical variables. Not all covariates in the selection models are given, and yet additional information such as math achievement scores and math private lessons are provided in Table 2. We found that school type by gender (*Gender.type*) and the area of school location (*City.size*) were common variables of which proportions were substantively different across classes within approaches. However, there were also model-specific variables whose means (or proportions)

Table 1 Consistency rates (%) in class membership across four models

Models	RE LOGIT		RE LINEAR		FE LOGIT		FE LINEAR	
(Class size)	Class 1 (28.2%)	Class 2 (71.8%)	Class I (48.7%)	Class II (51.3%)	Class A (79.6%)	Class B (20.4%)	Class X (43.6%)	Class Y (56.4%)
RE LOGIT			72.78		46.13		72.62	
RE LINEAR					53.83		68.25	
FE LOGIT							61.37	

Note LOGIT and LINEAR indicate latent class logistic and linear models, respectively. RE represents random effects, and FE represents fixed effects. Artificial number/characters are given for class names

varied depending on classes.[1] Among the variables, school-level covariates such as *Pct.disad* and *Dscpn* were different across classes in three approaches of the four. This implies that school-level covariates are important in explaining treatment selection heterogeneity. In addition, a few variables (*Sci.ach, Math.ach, Aca.emph*) were flagged as dissimilar across classes only in FE LOGIT. This indicates that the classes identified by FE LOGIT could differ from those estimated by the other models more qualitatively.

3.4 Covariate Balance Evaluation

Figure 1 displays covariate balance plots before and after balancing score adjustment for the first class of each model as well as for one homogeneous model. One homogeneous model assumed no subpopulations, and its PS was estimated via one class RE LOGIT. As seen from Fig. 1, one homogeneous model had less initial imbalanced covariates, and achieved equivalence between the treated and untreated groups in baseline covariates after MMS-W adjustment. For the two-class approaches, initial imbalanced covariates differed from class to class, and from model to model. After adjusting MMS-W, the four different approaches provided good covariate balance. FE LINEAR (far right bottom in Fig. 1) achieved covariate balance almost perfectly, while with FE LOGIT approach, one variable was still imbalanced in terms of its variance.

[1] In addition to *Gender.type* and *City.size*, model specific variables that differ between classes are *PL.sci*, *PL.math*, and *Dscpn* in RE LOGIT; *PL.sci*, *PL.math*, *Male, Dscpn,* and *Pct.disad* in RE LINEAR; *Sci.ach, Math.ach, Dad.edu, Pct.disad, Aca.emph,* and *Res.short* in FE LOGIT; *Sci.ach, Pct.disad, Res.short,* and *Dscpn* in FE LINEAR.

Table 2 Descriptive statistics (mean or proportion) of selection classes

Variables	RE LOGIT		RE LINEAR		FE LOGIT		FE LINEAR	
	Class 1	Class 2	Class I	Class II	Class A	Class B	Class X	Class Y
Level-1								
Sci.ach	557.70	555.99	559.14	553.93	554.63	563.66	554.32	558.13
Math.ach	608.25	605.11	609.88	602.30	603.79	614.58	603.13	608.21
Stu.conf.sci	8.61	8.64	8.72	8.56	8.62	8.68	8.62	8.65
Value.sci	8.93	8.95	8.99	8.89	8.93	9.00	8.88	8.98
PL.sci	0.42	0.29	0.42	0.23	0.33	0.31	0.37	0.28
PL.math	0.74	0.69	0.74	0.67	0.70	0.72	0.71	0.70
Male	0.52	0.49	0.55	0.45	0.49	0.52	0.49	0.50
Dad.edu								
_college	0.40	0.43	0.42	0.41	0.41	0.46	0.39	0.44
_noidea	0.32	0.31	0.31	0.31	0.31	0.29	0.30	0.31
Books25	0.86	0.86	0.86	0.86	0.86	0.87	0.85	0.87
Level-2								
Aca.emph	11.82	11.75	11.73	11.80	11.87	11.33	11.89	11.68
Res.short	11.17	11.10	11.14	11.10	11.07	11.31	11.04	11.17
Dscpn	10.66	11.16	10.71	11.31	11.06	10.92	11.06	11.02
Gender.type								
_Girlsch	0.16	0.22	0.13	0.26	0.21	0.17	0.23	0.18
_Coedu	0.63	0.59	0.65	0.56	0.62	0.55	0.57	0.62
Pct.disad								
_11–25%	0.32	0.36	0.36	0.34	0.34	0.38	0.43	0.30
_26–50%	0.24	0.25	0.22	0.28	0.28	0.10	0.28	0.23
_ > 50%	0.11	0.11	0.07	0.14	0.09	0.17	0.03	0.16
City.size								
_Urban	0.39	0.36	0.39	0.35	0.36	0.41	0.26	0.44
_Suburban	0.03	0.11	0.12	0.06	0.09	0.07	0.08	0.09
_Medium	0.34	0.27	0.32	0.26	0.28	0.31	0.34	0.25

Note Means are given for continuous variables, whereas proportions are provided for binary or categorical variables. Artificial numbers/characters are given for class names

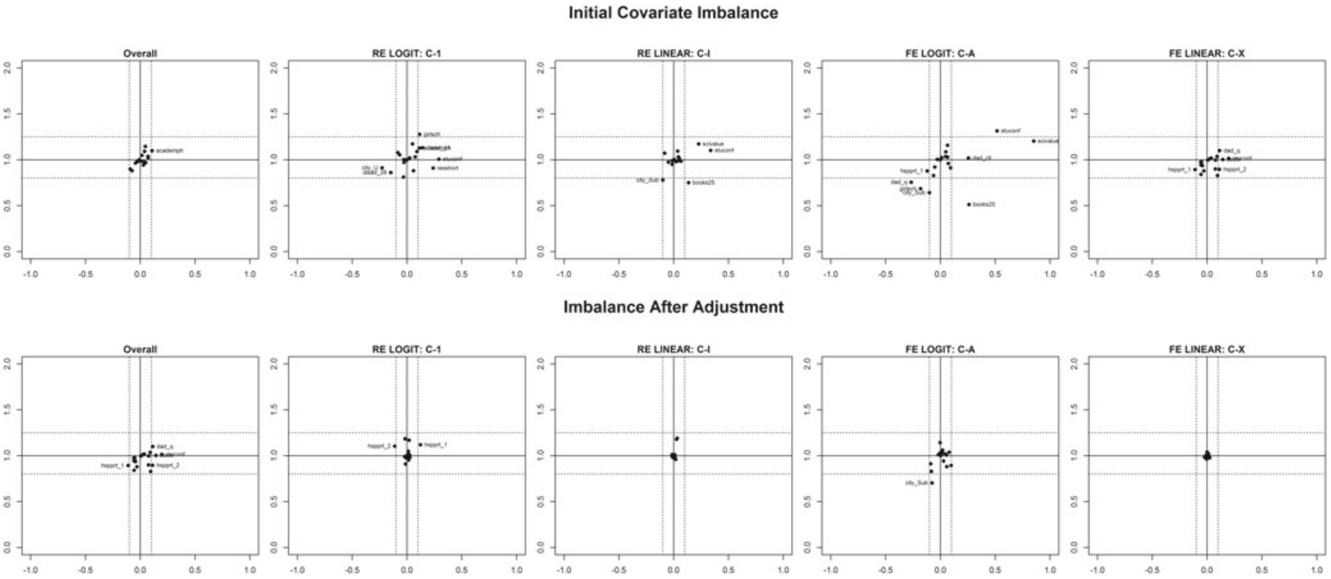


Fig. 1 Covariate balance plots before and after balancing score adjustment for one homogeneous class (overall) and the first classes of four multilevel mixture selection models

Table 3 ATE estimates of private science lessons

Estimates	RE LOGIT		RE LINEAR		FE LOGIT		FE LINEAR	
	Class 1	Class 2	Class I	Class II	Class A	Class B	Class X	Class Y
Prima facie (SE)	15.86* (4.13)	20.76* (2.83)	20.23* (3.14)	16.81* (3.62)	16.15* (2.58)	31.52* (5.24)	12.21* (3.37)	26.36* (3.21)
MMS-W (SE)	12.20* (4.05)	2.19* (2.76)	8.52* (3.06)	2.80* (3.51)	6.31* (2.52)	−1.69* (5.19)	7.05* (3.30)	4.64* (3.11)

Note LOGIT and LINEAR indicate latent class logistic and linear models, respectively. RE represents random effects, and FE represents fixed effects. SE indicates standard errors. MMS-W indicates marginal mean weighting through stratification

*Indicates that coefficients are statistically significant at the 0.05

3.5 ATE Within Classes

Prima facie effects are unadjusted mean differences between the treated and control groups. For the one homogeneous class model where we assume the homogenous selection processes across schools, the prima facie effect amounted to 19.10 points. After adjusting MMS-W, the effect decreased to 2.58 points, which was not significant. The one-class approach provided evidence that there was no effect of private science lessons overall.

Table 3 summarizes class-specific ATE estimates obtained by multilevel mixture modeling approaches. After estimating ATE within selection classes via four approaches, we found the unadjusted prima facie effects were significantly positive both in the first class of each approach and the second class of each approach. For example, the prima facie effects amounted to 15.86 and 20.76 for Classes 1 and 2, respectively in RE LOGIT. After MMW-S adjustment, the effect in the first class was still significantly positive, but reduced, whereas the adjusted effect in the second class was non-significant. Specifically, the adjusted effect of Class 1 was 12.20 points, which was significantly positive, whereas Class 2's effect of 2.19 points was not significant. Similar patterns were found in the other three approaches. These results imply that our within-class matching technique was very effective in practice to reveal distinctive effects for selection classes. Ignoring the existence of multiple selection classes may result in producing a misleading ATE that hides much of what goes on in subgroups.

However, the class size and class-specific treatment effects differed depending on the estimation models; when focusing on the first class of each model, the smallest class size for the first class, 28.2% in RE LOGIT, showed the largest estimate of 12.20 points, while the largest class size for the first class, 79.6% in FE LOGIT, showed the smallest estimate of 6.31 points. That is, we found that as the sample size in a positive-effects group increased, the corresponding ATE estimate decreased. Although the details on selection classes and the heterogeneity of ATE may be different depending on the chosen selection models, their pivotal conclusions did not disagree with each other: the existence of a positive effect group and no-effect group, resulting from heterogeneous selection mechanisms.

4 Discussion and Conclusions

We demonstrated four types of multilevel mixture models for identifying latent selection classes and estimating PS in a case study where the TIMSS dataset was used for examining the discovery of the heterogeneous effects of private science lessons. More specifically, class sizes, covariate balance, and treatment effect estimates were compared across different models using RE LOGIT as the baseline.

When using one homogeneous model that did not reflect the existence of multiple latent classes, we found no effect of private science lessons. Thus, we might conclude that private science lessons were not a useful resource for students to improve their science achievement scores on average. As the result shows, this across-cluster matching approach is not capable of discovering both selection mechanism heterogeneity and treatment effect heterogeneity. On the other hand, within-class matching with multilevel mixture selection models is effective in identifying homogeneous selection classes and naturally examining heterogeneous treatment effects for classes. Our multilevel mixture selection model approaches all reached the conclusion that some students got benefits from receiving private science lessons, while the others had no gains. We admitted that class sizes and class-specific estimates could vary depending on selection models, but treatment effect heterogeneity of each model could be explained by their distinctive motivation towards private science lessons of each.

In summary, this paper addresses selection model heterogeneity and treatment effect heterogeneity in PS analysis as a problem of subpopulations that may behave in selection processes differently and receive benefits of a treatment differently. The problem of selection model heterogeneity is closely associated with a problem of model specifications of latent class multilevel modeling. At present, the main approach to addressing the model specifications of latent class multilevel modeling is through exploring three alternatives to the conventional RE LOGIT. Thus, our paper accomplishes two goals. First, we provide an empirical example when within-class matching effectively worked. Again, the within-class matching retains the advantages of within-cluster and across-cluster matching and minimizes the disadvantages of both. This within-class matching is capable of explaining what is going on in subgroups more informatively. Second, we provide empirical evidence for Kim and Suk (2018)'s alternative model approaches of multilevel mixture selection models in PS analysis by addressing the problem of model specification for within-class matching. Using alternatives and comparing them with a baseline are of help to the estimation of heterogeneous treatment effects with multilevel assessment data, which is our goal. To conclude, we offer a multilevel mixture modeling approach in PS analysis that accounts for selection model heterogeneity and the corresponding treatment effect heterogeneity and that can be adopted by those who wish to do causal inference with multilevel assessment data.

Acknowledgements Support for this research was provided by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison with funding from the Wisconsin Alumni Research Foundation.

References

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311–359). Boston, MA: Springer.

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

Hong, G., & Hong, Y. (2009). Reading instruction time and homogeneous grouping in kindergarten: An application of marginal mean weighting through stratification. *Educational Evaluation and Policy Analysis, 31*(1), 54–81.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association, 101,* 901–910.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A, 171*(2), 481–502.

Kim, J. S., & Steiner, P. M. (2015). Multilevel propensity score methods for estimating causal effects: A latent class modeling strategy. In L. van der Ark, D. Bolt, W. C. Wang, J. Douglas, & S. M. Chow (Eds.), *Quantitative psychology research* (pp. 293–306)., Springer proceedings in mathematics & statistics Cham: Springer.

Kim, J.-S., Steiner, P. M., & Lim, W.-C. (2016). Mixture modeling strategies for causal inference with multilevel data. In J. R. Harring, L. M. Stapleton, & S. Natasha Beretvas (Eds.), *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications* (pp. 335–359). Charlotte, NC: IAP—Information Age Publishing, Inc.

Kim, J.-S., & Suk, Y. (2018, July). *Specifying multilevel mixture models in propensity score analysis*. Paper presented at the International Meeting of Psychometric Society, New York City, NY, US.

Leite, W. (2016). *Practical propensity score methods using R*. Sage Publications.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2016). *Methods and procedures in TIMSS 2015*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/publications/timss/2015-methods.html.

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

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

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

Steiner, P. M., & Cook, D. (2013). Matching and propensity scores. In T. Little (Ed.), *The oxford handbook of quantitative methods* (pp. 236–258). Oxford, England: Oxford University Press.

Suk, Y., & Kim, J.-S. (2018, April). *Linear probability models as alternatives to logistic regression models for multilevel propensity score analysis*. Paper presented at the American Educational Research Association, New York City, NY, US.

Wager, S. & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 1–15.

Specifying Multilevel Mixture Selection Models in Propensity Score Analysis

Jee-Seon Kim and Youmi Suk

Abstract Causal inference with observational data is challenging, as the assignment to treatment is often not random and people may have different reasons to receive or to be assigned to the treatment. Moreover, the analyst may not have access to all of the important variables and may face omitted variable bias as well as selection bias in nonexperimental studies. It is known that fixed effects models are robust against unobserved cluster variables while random effects models provide biased estimates of model parameters in the presence of omitted variables. This study further investigates the properties of fixed effects models as an alternative to the common random effects models for identifying and classifying subpopulations or "latent classes" when selection or outcome processes are heterogeneous. A recent study by Suk and Kim (2018) found that linear probability models outperform standard logistic selection models in terms of the extraction of the correct number of latent classes, and the authors continue to search for optimal model specifications of mixture selection models across different conditions, such as strong and weak selection, various numbers of clusters and cluster sizes. It is found that fixed-effects models outperform random effects models in terms of classifying units and estimating treatment effects when cluster size is small.

Keywords Causal inference · Finite mixture modeling · Latent class analysis · Selection bias · Balancing scores · Heterogeneous selection and treatment effects · Fixed-effects and Random-effects models · Hierarchical linear modeling

J.-S. Kim (✉) · Y. Suk
Department of Educational Psychology, Educational Sciences Building, University of Wisconsin-Madison, 1025 West Johnson Street, Madison, WI 53706, USA
e-mail: jeeseonkim@wisc.edu

Y. Suk
e-mail: ysuk@wisc.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_25

1 Heterogeneous Selection Processes

In multilevel observational data, treatment assignment is often not random, and treatment can be implemented at different levels (e.g., school-, class-, student-level, or time-varying treatments). Moreover, there may exist different reasons why some receive or are assigned to treatment, implying heterogeneity in treatment selection processes. Different motivations or selection criteria may result in different effects even if the received treatment is the same. It is thus important to account for the potential heterogeneity of the selection processes when making causal inferences with observational data.

Because unadjusted mean differences contain selection bias due to non-random assignment in addition to treatment effects, we need to remove selection bias to evaluate treatment effects properly. One of the most widely-used methods across various disciplines for making causal inference in observational studies is propensity score (PS) analysis (Rosenbaum & Rubin, 1983). A PS is a unit's conditional probability of receiving treatment Z_{ij} (1 = treated vs. 0 = untreated): $PS = \Pr\left(Z_{ij} = 1|\boldsymbol{X}_{ij}, \boldsymbol{W}_j\right)$ where $\boldsymbol{X}_{ij}$ and $\boldsymbol{W}_j$ are individual-level and cluster-level covariates, respectively, in two-level data where individual i is nested within cluster j. After a successful PS adjustment, the conditional distribution of baseline covariates would be similar for treated and untreated groups as in a randomized control trial.

Despite the common application of PS, removing selection bias is not straightforward and becomes more challenging when selection mechanisms differ across subgroups. Kim and colleagues (Kim & Steiner, 2015; Kim, Steiner, & Lim, 2016) presented a method to deal with heterogeneous selection processes by matching cases within homogeneous groups that share similar selection mechanisms. These homogeneous groups of individuals or clusters are referred to as "classes". The class membership might be known, for example, if the membership is related to measurable variables such as regions, districts, or the adoption of particular policies or practices. In those cases, the known class membership does not imply or require that the actual selection processes be known, as we can estimate selection processes for each class. For clustered data, multilevel logistic models are routinely used to estimate the selection process to account for the nesting and binary selection simultaneously. When class membership is unknown, the number and proportions of classes can be estimated using finite mixture modeling or latent class approaches (Clogg, 1995; McLachlan & Peel, 2000). Specifically, latent-class multilevel logistic models can be used as selection models to identify classes, classify members, and estimate PSs. Then treated and untreated cases are matched based on their estimated PSs. Finally, multilevel outcome models can be fitted for each class to estimate class-specific treatment effects (Kim & Steiner, 2015; Kim et al., 2016).

Logistic regression has predominantly been used for estimating selection models for multilevel logistic models where the random effects of clusters are natural extensions when data are clustered or nested. However, random-effects logistic models are not always optimal or superior to alternative approaches. Simulation studies on PS estimation with multilevel data revealed that both random-effects logistic models and fixed-effects logistic models are effective to estimate unbiased treatment effects (Arpino & Mealli, 2011; Leite et al., 2015; Thoemmes & West, 2011). Recently, Suk and Kim (2018) explored the potential of linear probability models as selection models and compared linear models with logistic models in multilevel PS analysis combined with heterogeneous treatment selection scenarios. They found that linear probability models outperformed logistic models in terms of identifying the correct number of latent classes, which is fundamental for estimating selection models and treatment effects. The current study extends this line of research by considering another type of model specification; that is, fixed-effects models for mixture selection models, and compares their properties with the commonly-used random effects models in multilevel data.

2 The Rubin Causal Model and Balancing Scores

The Rubin causal model (Rubin, 1974) was defined in the single-level data framework and has been extended to multilevel data (Hong & Raudenbush, 2006). In multilevel data, for example, $Y_{ij}(1)$ is the potential treatment outcome if individual i in cluster j is to be treated ($Z_{ij} = 1$), while $Y_{ij}(0)$ is the potential control outcome if untreated ($Z_{ij} = 0$). As only one of the two potential outcomes can be observed in reality, it is impossible to estimate an individual treatment effect τ_{ij}. However, the average treatment effect (ATE) can be defined as follows:

$$\tau_{ATE} = E(Y_{ij}(1)) - E(Y_{ij}(0)), \tag{1}$$

and this average effect can be estimated under the *strong ignorability assumption*, also called the *conditional independence assumption*, which is satisfied if all confounding covariates are measured validly and reliably, and if a conditional probability of being in the treatment group, given observed covariates, is strictly between zero and one (Rosenbaum & Rubin, 1983; Steiner & Cook, 2013). Under the strong ignorability assumption, potential outcomes are independent of treatment assignment after conditioning on all the confounding covariates:

$$\big(Y_{ij}(1), Y_{ij}(0)\big) \perp Z_{ij} | \boldsymbol{X}_{ij}, \boldsymbol{W}_j, \tag{2}$$

where $\boldsymbol{X}_{ij}$ and $\boldsymbol{W}_j$ are observed level-1 and level-2 covariates, respectively.

Rosenbaum and Rubin (1983) also introduced balancing scores in a single-level setting. A balancing score, $\boldsymbol{b}(\boldsymbol{X}_i)$, is a function of observed covariates $\boldsymbol{X}_i$ such that the conditional distribution of $\boldsymbol{X}_i$ given $\boldsymbol{b}(\boldsymbol{X}_i)$ is identical for treated and untreated groups; that is,

$$\boldsymbol{X}_i \perp Z_i | \boldsymbol{b}(\boldsymbol{X}_i). \tag{3}$$

Balancing scores can be any scores satisfying the condition in Eq. (3); they can be covariates themselves (the finest balancing scores) or a PS (the coarsest). The PS is the conditional probability that an individual receives the treatment and is bounded below by zero and above by one (i.e., 0 < PS < 1). However, we use linear models as alternatives to logistic models, and they can produce predicted values outside the [0, 1] range, so we use balancing scores, more general forms than propensity scores, for our linear selection models. Rosenbaum and Rubin's balancing scores can be extended to multilevel data:

$$\boldsymbol{X}_{ij}\boldsymbol{W}_j \perp Z_{ij} | \boldsymbol{b}\left(\boldsymbol{X}_{ij}, \boldsymbol{W}_j\right),$$

and we use this multilevel version of balancing scores for linear selection models.

3 Mixture Selection Models Specifications

Multilevel PS matching techniques can be explained using a multilevel matching continuum (Kim et al., 2016). The two extremes of the continuum are within-cluster matching and across-cluster matching. Within-cluster matching finds matches between treated and control units for each cluster separately, only using individual-level covariates, whereas across-cluster matching finds matches for all clusters at once. Although within-cluster matching has a number of important theoretical advantages, such as no need to balance on cluster-level covariates and weaker identification assumptions, the idea of local matching within each cluster is not always plausible for real data due to the lack of overlap for some or most clusters in practice, particularly with small cluster sample sizes or strong treatment selection.

To overcome insufficient overlap within clusters, one can "borrow" treated or control units from other clusters, and this is the concept of across-cluster matching. In across-cluster matching, both individual-level and cluster-level covariates affect the selection process, and thus covariates at both levels should be measured reliably and specified correctly in the selection model. Across-cluster matching is proper and effective when clusters are similar to each other and have homogeneous selection processes. However, when distinctive selection processes exist across clusters, implementing across-matching implies the violation of the strong ignorability assumption and can be misleading and detrimental (Kim, Lubanski, & Steiner, 2018).

Kim and Steiner (2015) presented a new matching technique by pooling units within homogeneous groups of clusters or classes. If membership is known, matches can be made by pooling units within manifest classes (e.g., types, divisions, sectors), and if membership is unknown, homogeneous latent classes can be found using mixture modeling. Homogeneous classes, ether known or estimated, can be defined in both selection and outcome models. Compared to within-cluster matching, this matching strategy is beneficial in that we can obtain larger samples for the homogeneous groups, thus increasing the overlap between the treated and control units (Kim et al., 2016).

Suk and Kim (2018) explored alternatives to logistic models as a multilevel mixture selection model. They found that linear probability models extracted the number of latent classes more accurately than logistic models. As an extension of this research, we proposed using fixed effects (FE) models for specifying level-2 models. It is well-known that random effects (RE) are frequently incorporated into various forms of regression models, to account for correlated responses within clusters and multiple sources of variance. RE is efficient and frequently preferred over FE. However, the use of RE is valid if the underlying assumptions hold; one of the assumptions is that the cluster-specific effects are not correlated with independent variables in the model. If the RE assumption is satisfied, RE models are more efficient, whereas when not satisfied, RE models are not consistent (Cameron & Trivedi, 2005; Nerlove, 2005). Since Hausman's model specification test in 1978, FE has been compared to RE repeatedly (Gui, Meierer, & Algesheimer, 2017; Hausman & Taylor, 1981; Kim & Frees, 2007), and this comparison is grounded on the fact that RE is more efficient while FE is more consistent. In PS analysis, since prediction accuracy of selection models is a priority over efficiency, it would be valuable to explore the potential of FE in PS analysis, particularly when selection scenarios are not simple (e.g., heterogeneous selection processes). Therefore, this paper investigated the performance of the two alternatives—RE linear and FE logistic models—for mixture selection models with simulated multilevel data, by comparing them with RE logistic models as a baseline.

Our investigation of FE models is rooted in their desirable properties against omitted variable bias. Hausman (1978) compared FE with RE models for panel data and developed an omitted variable test between the robust FE and efficient RE estimators. Hausman's test has been used and modified for the past 40 years (Gui et al., 2017; Hausman & Taylor, 1981; Kim & Frees, 2007). The comparison is based on the theoretical properties of the RE estimator being consistent and most efficient in the absence of omitted confounders but unfortunately sensitive to omitted variables. On the other hand, the FE estimator is robust against time-constant omitted confounders but is inefficient compared to the RE estimator and thus not always preferred, especially when model parsimony and efficiency is desired.

As prediction accuracy is a critically important quality for selection models, while efficiency and parsimony is usually not, it is reasonable and seems natural to use FE selection models, especially when selection mechanisms are complex and heterogeneous. Therefore, this study examines and compares three model specifications of mixture selection models with multilevel observational data: the common RE logistic modeling is considered as the baseline, and Suk and Kim's RE linear probability modeling and FE logistic modeling as two alternatives to RE logistic modeling. For FE logistic modeling, latent-class logistic models with dummy variables for clusters are used.

4 Simulation Design and Data Generating Model

Simulation data were generated with two selection classes in multilevel structures. In the simulation design, a random sample of J clusters with $N = n_1 + n_2 + \cdots + n_J$ individual units was generated based on RE linear models for the outcome model with a continuous outcome Y_{ij} and RE logistic models for the selection model with a binary treatment Z_{ij}. Both models included two level-1 covariates (X_1, X_2) and two level-2 covariates (W_1, W_2). Level-1 covariates depended on level-2 covariates in the underlying model. The first selection class, Class 1, had non-zero effects of covariates in the selection model and positive effects of the treatment in the outcome model. The second selection class, Class 2, had no effects of covariates in the selection model and no treatment effects in the outcome model. Each unit's treatment status ($\mathrm{Z}_{ijs} = 1$ or 0) was randomly assigned from a Bernoulli distribution with the selection probability $\pi_{ijs}, \mathrm{Z}_{ijs} \sim Bernoulli(\pi_{ijs})$. The true selection probability was generated through the following RE logistic regression model:

$$\mathrm{logit}(\pi_{ijs}) = \alpha_s + X'_{ijs}\beta_{js} + W'_{js}\gamma_s + T_{js} \tag{4}$$

where $\mathrm{i} = 1, \ldots, n_{js}$, $\mathrm{j} = 1, \ldots, M_s$, $\mathrm{s} = 1, 2$ denote level-1 unit, cluster, and class, respectively; π_{ijs} is the propensity of receiving a treatment for a level-1 unit i in cluster j in class s; X_{ijs} and W_{js} are level-1 covariates and level-2 covariates, respectively; T_{js} are random effects for cluster j; α_s are class-specific intercepts; β_{js} and γ_s are class-specific level-1 regression coefficients and level-2 regression coefficients, respectively.

The simulation study was conducted with five factors hypothesized to influence the performance of the estimators: (1) the estimation models (random-effects logistic, random-effects linear, fixed-effects logistic models; RE LOGIT, RE LINEAR, FE LOGIT), (2) class size (equal, unequal), (3) number of clusters (small, large), (4) cluster size (small, medium, large) and (5) strength of selection (strong, moderate, weak).

The first factor was the estimation model specification with RE LOGIT, RE LINEAR, and FE LOGIT as the three levels of the factor. The first model RE LOGIT is the data generating model in Eq. (4). The model specification of RE LINEAR is as follows:

$$Z_{ijs} = \alpha_s + X'_{ijs}\beta_{js} + W'_{js}\gamma_s + T_{js} + \epsilon_{ijs} \tag{5}$$

where Z_{ijs} is the treatment assignment variable for a level-1 unit i in cluster j in class s, $Z_{ijs} \in 0, 1$ (0 = untreated; 1 = treated); ϵ_{ijs} is the individual-specific random effect for a level-1 unit i in cluster j in class s, and the other notations are the same as in Eq. (4).

FE LOGIT is specified as follows:

$$\mathrm{logit}(\pi_{ijs}) = \alpha_s + X'_{ijs}\beta_{js} + D'_{js}\gamma_s + T_{js} \tag{6}$$

where D_{js} is a dummy variable for each cluster j > 1 (omitting one of cluster dummies because of multicollinearity), and there are no level-2 covariates, W_{js}. The other notations are the same as in Eq. (4).

We identified latent classes using the Mplus8 program (Muthén & Muthén, 1998–2017) and used a "class assignment-based" algorithm to solve the label switching problem (Tueller, Drotar, & Lubke, 2011). We obtained propensity and balancing scores estimated by RE LOGIT, RE LINEAR, and FE LOGIT using the R program (R Core Team, 2017) with the lme4 package as follows:

```
RE.LOGIT.ps_C1 <- predict(glmer(Z ~ 1 + X1 + X2 + W1 + W2 + (1|clusterid),
  data = df.cg1, family = binomial), type = 'response') # PS of Class 1 via
  RE LOGIT

RE.LINEAR.bs_C1 <- predict(lmer(Z ~ 1 + X1 + X2 + W1 + W2 + (1|clusterid),
  data = df.cp1)) # BS of Class 1 via RE LINEAR

FE.LOGIT.ps_C1 <- predict(glm(Z ~ 1 + X1 + X2 + id, data = df.fecg1, family
  = binomial), type = 'response') # PS of Class 1 via FE LOGIT
```

In addition, we obtained the treatment effects before and after balancing score adjustment, respectively. Particularly, as our balancing score adjustment, we applied Hong and Hong (2009)'s marginal mean weighting through stratification.

The second factor was class size, considered at equal and unequal levels. For the equal condition, the sample size proportions of Class 1, *nC1*, and Class 2, *nC2*, were 50% versus 50%, while for the unequal condition, the proportions were 70% versus 30%. The third factor was the number of clusters, nC ($= nC1 + nC2$), considered at levels of 50 and 100. The fourth factor was cluster size, which is the number of level-1 units, *nI*. It had three levels of 20, 30, and 50 where each level follows a normal distribution with varying mean and standard deviation: N(20, 3), N(30, 5), and N(50, 10). The fifth factor was the strength of selection, considered at levels of strong, moderate, and weak levels; strong selection had 70–75% average within-cluster overlap, while moderate selection had 75–80% and weak selection had 80–85% average within-cluster overlap, respectively.

The performance of the ATE estimates was evaluated with respect to the absolute value of the remaining bias, simulation standard deviation (SD) and mean squared error (MSE). For R simulation replications ($r = 1,\ldots,$ R), the three criteria with estimated treatment effect $\hat{\theta}_r$ and true treatment effect τ are defined as follows:

$$|\mathrm{Bias}| = \left| \frac{1}{R} \sum_{r=1}^{R} \frac{\hat{\theta}_r - \tau}{\tau} \right|,$$

$$\mathrm{SD} = \sqrt{\frac{1}{R-1} \sum_{r=1}^{R} \left(\hat{\theta}_r - \bar{\hat{\theta}}_r \right)^2},$$

and

$$\mathrm{MSE} = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_r - \tau\right)^2.$$

When the true treatment effect is zero ($\tau = 0$), bias is defined by the average of the estimated treatment effects $\hat{\theta}_r$ across the replications.

5 Simulation Results

5.1 Extraction of Latent Classes

Figure 1 summarizes extraction results under two different selection processes (Classes 1 & 2) using the RE LOGIT, RE LINEAR, and FE LOGIT approaches. The proportions (%) of correct extraction were calculated on the basis of Akaike information criterion (AIC) with 200 replications. There were two types of incorrect extraction: over-extraction and under-extraction. Over-extraction represents when the model fit indexes favored models with more than two classes. Under-extraction represents when the model fit indexes favored the one class model. The three panels in Fig. 1 depict the effects of the specification of estimation model (RE LOGIT, RE LINEAR, FE LOGIT), cluster sizes (20, 30, 50), and selection strength (Strong, Moderate, Weak). Separate plotting points in the same color for the Strong, Moderate, and Weak selection conditions within the panels represent the four combinations of the number of clusters between two classes for each estimation model; (*nC1*, *nC2*) = (25, 25), (35, 15), (50, 50), and (70, 30). The effects of the cluster sizes are found in the expected direction (larger performs better), but sometimes the differences were minimal and the results overlapped in these conditions.

One of the major findings from the simulation is that the estimation model is the most important factor influencing the extraction of the correct number of latent classes; RE LINEAR outperformed the baseline RE LOGIT in terms of extraction across the simulation conditions in the study. FE LOGIT showed under-extraction in most conditions, especially with small cluster sizes or weak selection. Only when the cluster size is 50 and the selection is strong, FE LOGIT extracted the correct number of latent classes. We also found that the proportions of correct extraction were generally higher in the strong selection condition than in the weak selection condition, but the effects of the strength of selection and cluster sizes were not always consistent, suggesting moderating effects among the five factors. Whereas FE LOGIT frequently under-extracted the number of latent classes, RE LOGIT was incorrect in the opposite direction and showed a tendency to over-extract the number of latent classes. Therefore, among the three estimation models, RE LINEAR appears to be the best option regarding class extraction and shows more consistent performance than the other models with over 80% correct extraction in many and over 90% in some conditions. The proportion of correct extraction for RE LINEAR was as low as only 50% when selection is weak, the number of individuals within cluster is small

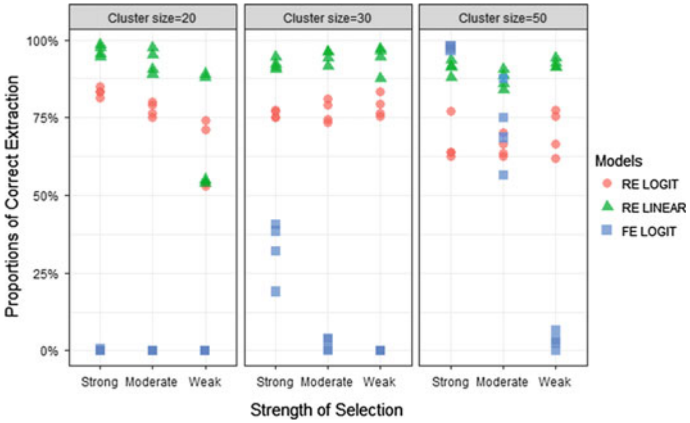


Fig. 1 Extraction accuracy in various combinations of estimation models, selection strength, and cluster sizes

(cluster size = nI = 20), and the number of clusters is no more than 25 per class and 50 in total ($nC = nC1 + nC2 = 25 + 25 = 50$). Even in this condition, RE LINEAR performed better than RE LOGIT and FE LOGIT, indicating that correct class extraction is difficult when selection is weak and cluster size is small.

We also found that, cluster size (# level-1 units) and the number of clusters (# level-2 units) have an interaction effect on class extraction; while the two factors do not affect each other clearly when the number of clusters is relatively small ($nC = 50$), large cluster sizes result in over-extraction of latent classes when the number of clusters is large ($nC = 100$). Although over-extraction of latent classes with large samples is not an ideal character of selection models, having multiple latent classes with similar selection processes is generally less problematic than an inability to identify distinctive selection processes with respect to understanding causal mechanism.

5.2 *Classification of Units into Latent Classes*

Figure 2 summarizes the proportions (%) of correct classification with 100 replications, where correct classification implies cluster membership was accurately estimated using latent class models by the highest posterior probabilities. As a correct classification cannot be made with an incorrect number of latent classes, we only considered cases where the extraction of two latent classes was achieved first.

We found that the differences among the three estimation models were relatively small with respect to classification compared to the large variation in class extraction in Fig. 1. Although the main effect of the estimation models was insignificant, we found interesting interaction effects between the models and the other factors. Specif-

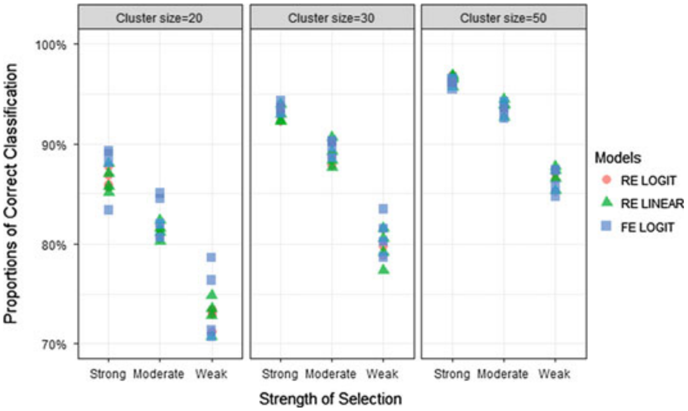


Fig. 2 Classification accuracy in various combinations of estimation models, selection strength, and cluster sizes

ically, FE LOGIT was more accurate than the baseline RE LOGIT in classification when class sizes were unequal (70% vs. 30%). Moreover, the difference between FE LOGIT and RE LOGIT with unequal class sizes became greater in smaller cluster sizes, indicating the choice of the estimation model is especially important when the sizes of subpopulations are not similar, and cluster sizes are small.

Clear patterns in Fig. 2 demonstrate that correct classification is largely affected by the strength of selection and cluster sizes. Classification was most accurate, over 95%, with strong selection and large cluster size (50). Proportions of correct classification were 80% or higher across conditions except when the selection was weak and cluster size is not large (20 or 30). It is understandable that classification is more difficult when selection is weak and the characteristics of the two classes are rather similar, as in a mixture normal distribution with a large overlap in the middle. In this case, identifying two heterogeneous subpopulations is not as critical compared to two classes with vastly different characteristics, supporting this ideal property of latent class models.

5.3 Estimation of Average Treatment Effects

Figure 3 summarizes the results of ATE estimation after balancing score adjustment in Class 1, where the selection is not random and the true treatment effect is positive. To make the comparisons across the conditions interpretable, the results are based on 100 replications where the number of latent classes, two, is correctly identified. As expected, balancing score adjustment removed most bias when the cluster size is large, selection is weak, and the number of clusters is large. In comparing the three estimation models, RE LINEAR performed similarly to or better than the baseline

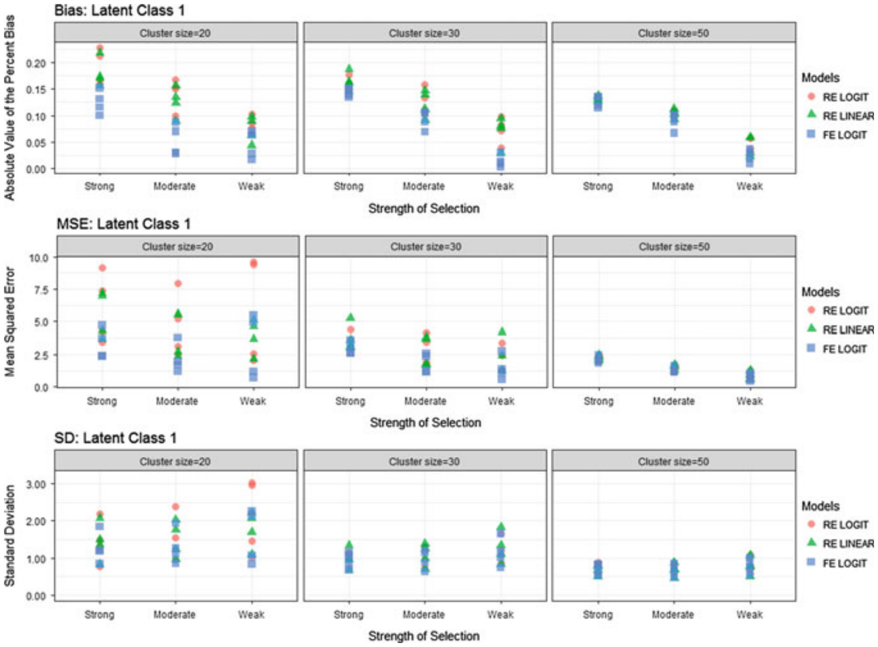


Fig. 3 The bias, SD, and MSE of the average treatment effect (ATE) estimates in various combinations of estimation models, selection strength, and cluster sizes

RE LOGIT in terms of bias removal across the simulation conditions. It is noteworthy that FE LOGIT removed more bias than RE LOGIT in most conditions, particularly so in the (1) small cluster size and (2) large cluster size combined with weak selection conditions.

The SD and MSE decreased as sample size increased at varying degrees. The biggest difference was found in the cluster sizes between 20 and 30, while the difference was rather small between 30 and 50. The number of clusters also affected SD and MSE, and RE LOGIT showed much larger MSE than the others when the number of clusters and cluster size were both small. FE LOGIT had smallest bias, SD, and MSE in most conditions. The effects of class proportions and the number of clusters were found to be minimal when the cluster size reached 50. These results imply that RE LOGIT needs a large sample size to obtain unbiased ATE estimates with high precision, compared to the other models we examined, RE LINEAR and FE LOGIT. With small numbers of clusters and individuals in particular, FE LOGIT can be an effective alternative to the standard RE LOGIT in removing selection bias.

6 Conclusions

We investigated the properties of three model specifications for multilevel mixture selection models and compared their relative performances across plausible settings in practice. Although random-effects logistic models (RE LOGIT) are routinely used in multilevel PS analysis, their popularity does not guarantee that RE LOGIT is an optimal choice. Our simulation study supports the main finding by Suk and Kim (2018) in that RE LOGIT tends to overestimate the number of latent classes, and linear probability models at level 1 (RE LINEAR) exhibit greater consistency and accuracy in terms of class extraction than RE LOGIT. Incorrect identification of latent classes is a critical issue for within-class matching strategies, because the further analyses of PS adjustment and ATE estimation are greatly affected by class extraction.

This study also emphasizes the properties of fixed-effects specification of clusters at level 2 (FE LOGIT) and found that FE LOGIT classifies units correctly more often than RE LOGIT when class sizes were unequal and cluster sizes are small. Differences among RE LOGIT, RE LINEAR, and FE LOGIT were small with large cluster sizes. The choice of the selection model also affects the ATE estimation, the ultimate goal of PSA. We found that RE LOGIT requires large sample sizes to estimate unbiased ATE with small MSE, and is more sensitive to small sample sizes than the other models. On the other hand, FE LOGIT showed small remaining bias and MSE with small numbers of clusters and cluster sizes.

In conclusion, although RE LOGIT has been used as a natural extension of logistic regression for multilevel data, we can consider other approaches given the specific conditions of the data, such as heterogeneous selection or outcome processes, known or unknown homogenous group memberships, relative sizes of latent classes, strength of selection, the number of clusters, and cluster sizes. We can also use different models for the different steps of PSA; for example, RE LINEAR for class extraction and FE LOGIT for classification of units as different specifications have specific strengths. Finally, for real data analysis where the true selection mechanism is unknown, we can implement several specifications of models to compare their results, evaluate the validity of assumptions, and strengthen our inferences. For an empirical example of applying different multilevel mixture selection models to identify potentially heterogeneous ATEs, we refer to Suk and Kim in this volume.

References

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis, 55,* 1770–1780.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. Cambridge: Cambridge University Press.

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311–359). Boston, MA: Springer.

Gui, R., Meierer, M., & Algesheimer, R. (2017). REndo: Fitting linear models with endogenous regressors using latent instrumental variables. R package version 1.3. https://CRAN.R-project.org/package=REndo.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica, 46,* 1251–1271.

Hausman, J. A., & Taylor, W. E. (1981). Panel data and unobservable individual effects. *Econometrica, 49,* 1377–1398.

Hong, G., & Hong, Y. (2009). Reading instruction time and homogeneous grouping in kindergarten: An application of marginal mean weighting through stratification. *Educational Evaluation and Policy Analysis, 31,* 54–81.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association, 101,* 901–910.

Kim, J. S., & Frees, E. W. (2007). Multilevel modeling with correlated effects. *Psychometrika, 72,* 505–533.

Kim, Y., Lubanski, S. A., & Steiner, P. M. (2018). Matching strategies for causal inference with observational data in education. In C. Lochmiller (Ed.), *Complementary research methods for educational leadership and policy studies* (pp. 173–191). Cham: Palgrave Macmillan.

Kim, J. S., & Steiner, P. M. (2015). Multilevel propensity score methods for estimating causal effects: A latent class modeling strategy. In L. van der Ark, D. Bolt, W. C. Wang, J. Douglas, & S. M. Chow (Eds.), *Quantitative psychology research* (pp. 293–306). Cham: Springer.

Kim, J.-S., Steiner, P. M. & Lim, W.-C. (2016). Mixture modeling strategies for causal inference with multilevel data. In J. R. Harring, L. M. Stapleton, & S. Natasha Beretvas (Eds.), *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications* (pp. 335–359). Charlotte, NC: IAP—Information Age Publishing, Inc.

Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., MacInnes, J. W., & Sandbach, R. (2015). An evaluation of weighting methods based on propensity scores to reduce selection bias in multilevel observational studies. *Multivariate Behavioral Research, 50,* 265–284.

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

Muthén, L. K., Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Nerlove, M. (2005). *Essays in panel data econometrics*. Cambridge: Cambridge University Press.

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

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

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701.

Steiner, P. M., & Cook, D. (2013). Matching and propensity scores. In T. Little (Ed.), *The oxford handbook of quantitative methods* (pp. 236–258). Oxford, England: Oxford University Press.

Suk, Y., Kim, J.-S. (2018, April). *Linear probability models as alternatives to logistic regression models for multilevel propensity score analysis*. Paper presented at the annual meeting of American Educational Research Association, New York City, NY.

Thoemmes, F. J., & West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research, 46,* 514–543.

Tueller, S. J., Drotar, S., & Lubke, G. H. (2011). Addressing the problem of switched class labels in latent variable mixture model simulation studies. *Structural Equation Modeling, 18,* 110–131.

The Effect of Using Principal Components to Create Plausible Values

Tom Benton

Abstract In all large scale educational surveys such as PISA and TIMSS the distribution of student abilities is estimated using the method of plausible values. This method treats student abilities within each country as missing variables that should be imputed based upon both student responses to cognitive items and a conditioning model using background information from questionnaires. Previous research has shown that, in contrast to creating single estimates of ability for each individual student, this technique will lead to unbiased population parameters in any subsequent analyses, provided the conditioning model is correctly specified (Wu in Studies in Educational Evaluation 31:114–128, 2005). More recent research has shown that, even if the conditioning model is incorrectly specified, the approach will provide a good approximation to population parameters as long as sufficient cognitive items are answered by each student (Marsman, Maris, Bechger, & Glas in Psychometrika 81:274–289, 2016). However, given the very large amount of background information collected in studies such as PISA, background variables are not all individually included in the conditioning model, and a smaller number of principal components are used instead. Furthermore, since no individual student answers cognitive items from every dimension of ability, we cannot rely on sufficient items having been answered to ignore possible resulting misspecification in the conditioning model. This article uses a simple simulation to illustrate how relying upon principal components within the conditioning model could potentially lead to bias in later estimates. A real example of this issue is provided based upon analysis of regional differences in performance in PISA 2015 within the UK.

Keywords Plausible values · Principal components · PISA

T. Benton (✉)
Cambridge Assessment, The Triangle Building, Shaftesbury Road, Cambridge CB2 8EA, UK
e-mail: benton.t@cambridgeassessment.org.uk

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_26

1 Introduction

The Programme for International Student Assessment (PISA) is an international survey run once every three years since 2000 by the Organization for Economic Co-operation and Development (OECD). It aims to evaluate education systems around the world by testing the skills and knowledge of 15-year-old students. Detailed information on the methodology used within the latest PISA study (PISA 2015) is given in OECD (2017). This paper will focus upon one particularly part of the methodology—the use of plausible values.

Within each PISA study, across a number of different countries, samples of students attempt to answer sets of items testing their skills in math, reading and science. This allows a comparison of student abilities on a common basis across different countries. In the PISA data sets, estimates of students' abilities, based on their responses to the cognitive items, are recorded in the form of plausible values.

Plausible values are used within PISA, as well as other international surveys such as PIRLS and TIMSS, to overcome the ubiquitous problem of measurement error in educational assessments (Laukaityte & Wiberg, 2017). Measurement error in this context refers to the fact that the performance of individuals can vary depending upon the precise selection of items to which they are asked to respond. Whilst this type of variation is acknowledged in educational research, often it is not formally accounted for. In contrast, within international surveys, through the use of plausible values, all published estimates of the distributions of abilities within different subgroups of students are designed to have fully accounted for the possible influence of measurement error. Failing to account for measurement error can lead to underestimating the size of the differences between different subgroups (Von Davier, Gonzalez, & Mislevy, 2009). With this and several other technical considerations in mind, it has been asserted that the use of plausible values remains "the state-of-the-art for secondary analyses" of the PISA databases (Braun & Von Davier, 2017).

The motivation for this research came from finding a clash between the optimism of the technical research cited above and some confusing results found in practice during secondary analysis of the PISA 2015 data in the UK. The beginnings of this concern are illustrated in Fig. 1. This figure, which was produced by collating information from Jerrim and Shure (2016a) and Jerrim and Shure (2016b), shows how the difference between top performers in reading (i.e., those at the 90th percentile) in Northern Ireland and England (two separate parts of the UK) has changed over time.

As can be seen, Fig. 1 indicates that, whilst from 2006 to 2012 there was little difference in the scores of top performers between England and Northern Ireland, in 2015 a large gap appeared. This could be taken as an indicator that Northern Ireland is no longer getting the best performance out of its most able students. Similar results were found for both math and science (Jerrim & Shure, 2016b, pp. 52 and 88). Although not shown here, the published results also seemed to indicate that amongst the lowest performers (i.e., those at the 10th percentile) Northern Ireland's scores

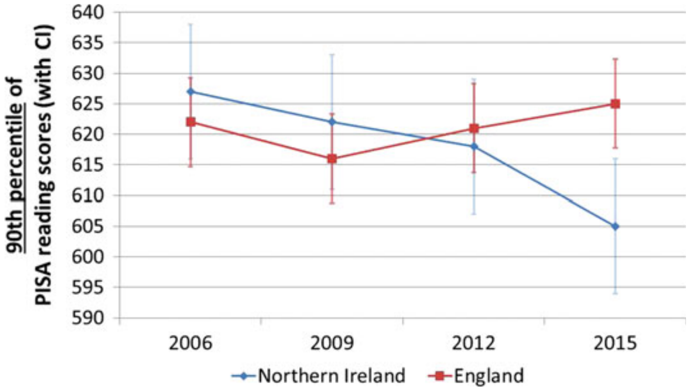


Fig. 1 The 90th percentile of reading scores in England and Northern Ireland between 2006 and 2015. Error bars indicate 95% confidence intervals for each estimate

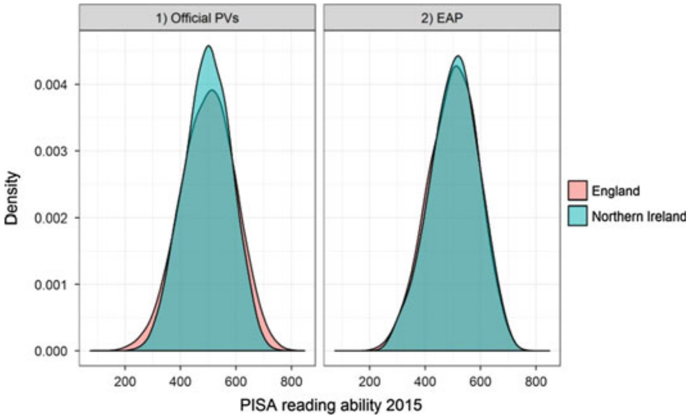


Fig. 2 Comparing distributions of reading ability between England and Northern Ireland using the official plausible values and using unidimensional EAP estimates

had substantially increased between 2009 and 2015 (Jerrim & Shure, 2016b pp. 55, 91 and 107).

Another way to illustrate this same issue is in Fig. 2. As can be seen, the official PISA plausible values (the left-hand side) suggested that the distribution of reading ability is narrower in Northern Ireland than it is in England. However, if single ability estimates are calculated for each student based purely upon their performance on the reading items (EAP estimates), no major difference in the spread of ability can be seen between England and Northern Ireland.

Of course, these results do not prove that the official plausible values are incorrect. As discussed, existing technical literature provides good reasons why plausible values are used rather than simple individual estimates of ability for each student. Nonetheless, it was of interest to investigate the source of the discrepancy.

The aim of this paper is to investigate how differences in results derived from plausible values and those derived from point estimates of student abilities, such as those shown above, might occur in practice. The paper will also show how the detailed method used to create plausible values, and the use of principal components analysis as a data reduction technique within this process in particular, could potentially lead to bias in parameter estimates for particular subgroups of students. For example, it will show that the methodology could potentially produce the impression of a reduced spread of abilities within Northern Ireland even if, in fact, none existed.

2 How Are Plausible Values Produced and Why May It Be a Problem?

In almost all psychometric work, it is realized that a student's ability in a particular domain is a latent trait that cannot be directly observed. That is, ability is an unobserved missing variable. What is observed is each student's performance in a particular assessment. We use these observations to infer the likely location of their ability.

At this point there are two approaches. Often, analysts create a single estimate of each student's ability on some scale. For example, in simple cases this might be achieved by adding up their scores from each item in a test. In other cases, such as computer adaptive tests, where different students attempt different sets of items, item response theory (IRT) or Rasch analysis might be used to produce scores. Having done this, the fact that the students might have achieved slightly differently given a different set of items to answer (i.e., measurement error) is simply acknowledged as a caveat and results reporting continues. In contrast, plausible values handle this missing data problem explicitly through the process of multiple imputation. Rather than assigning a single value to each student, several values are imputed from amongst the likely range of actual abilities.

In the international surveys, plausible value imputation is done using a combination of two elements: a multidimensional IRT model, and a conditioning model. The IRT element captures the fact that the greater a student's ability in each domain, the greater their chances of answering items correctly. As such, information from item scores can be used within the imputation process; students who have answered a greater number of items correctly are likely to have higher ability values imputed. Alongside this, a conditioning model uses latent regression to estimate the relationship between a range of background factors and ability within each participating country. The coefficients of this model are used to inform the imputation process to ensure that later estimates of the relationship between background variables (e.g., student gender) and abilities are unbiased. More details of this process are given in Wu (2005).

Very briefly, and following the notation of Wu (2005), the procedure is as follows. First, we denote the vector of item responses to M items in a test for the ith student as

$\boldsymbol{x}_i = \{x_{1i}, \ldots, x_{Mi}\}$. The vector of abilities associated with this student is denoted as $\boldsymbol{\theta}_i$. Note that if we are dealing with a unidimensional test, then $\boldsymbol{\theta}_i$ will be a single number whereas, if it is multidimensional, it will be a vector. The probability of the student's set of item responses is given by $f(\boldsymbol{x}_i|\boldsymbol{\theta}_i) = \prod_m f_m(x_{mi}|\boldsymbol{\theta}_i)$. In the case of PISA 2015, each f_m is defined according to the generalized partial credit model (see OECD, 2015, p. 143). We denote the vector of background information about each student, containing information such as indicators of the school they attend, and their responses to questions in the student questionnaire, as $\boldsymbol{y}_i$. The conditional distribution of $\boldsymbol{\theta}_i$ dependent upon the background data is given by the following formula.

$$g(\boldsymbol{\theta}_i|\boldsymbol{y}_i) \sim N\left(\boldsymbol{y}_i^T\Gamma, \Sigma\right) \tag{1}$$

In Eq. (1), Γ denotes a matrix of latent regression coefficients, and Σ denotes the joint residual variance matrix of $\boldsymbol{\theta}_i$. If we are dealing with a unidimensional test, then Σ will simply be a single number (usually denoted σ^2). The number of columns of Γ will be equal to the number of dimensions of the test being analyzed.

The item parameters in each item response probability function f_m, and the latent regression parameters Γ and Σ can be estimated using a maximum likelihood method. Once this has been done, the posterior distribution of each student's $\boldsymbol{\theta}_i$ vector is given by the equation below.

$$\mathrm{h}(\boldsymbol{\theta}_i|\boldsymbol{y}_i, \boldsymbol{x}_i) = \frac{\mathrm{f}(\boldsymbol{x}_i|\boldsymbol{\theta}_i)\mathrm{g}(\boldsymbol{\theta}_i|\boldsymbol{y}_i)}{\int \mathrm{f}(\boldsymbol{x}_i|\boldsymbol{\theta}_i)\mathrm{g}(\boldsymbol{\theta}_i|\boldsymbol{y}_i)d\boldsymbol{\theta}_i} \tag{2}$$

Rather than calculating a single estimate of $\boldsymbol{\theta}_i$ for each student, the plausible values methodology samples several values (typically 5 or 10) from this posterior distribution. In some cases, the posterior distribution may be approximated by a multivariate normal distribution to simplify the process of sampling plausible values (OECD, 2017, p. 147). Procedures for producing plausible values in this way are now included as a standard feature of many IRT software packages including the R packages TAM (Robitzsch, Kiefer, & Wu, 2017) and mirt (Chalmers, 2012).

In the case of PISA 2015, the conditioning model aims to include a huge number of background variables that are collected in accompanying questionnaires. Furthermore, before use in the conditioning model each of these variables is contrast coded so that, for example, rather than including a single variable for student age, 14 variables are created denoting whether each student is below 15.25 years of age, whether they are between 15.25 and 15.33, whether they are between 15.33 and 15.42 and so on. In fact, if information from all PISA questionnaires is available within a country, around 3000 indicators are included in the conditioning model.

Because, for most countries, the number of variables included in such a model would actually be of a similar order of magnitude to the number of sampled students in the country, the resulting coefficients would be highly unstable if the conditioning model were fitted directly. To address this, before applying the conditioning model, the thousands of indicators are condensed using principal components analysis (PCA). This classical statistical technique is designed to retain as much information as possible from the original data but using a much smaller number of variables. That is, the full vector of background characteristics ($\boldsymbol{y}_i$) in Eqs. (1) and (2) is replaced by a much shorter vector of principal component scores ($\boldsymbol{t}_i$) and the conditional distribution of $\boldsymbol{\theta}_i$ dependent upon the background information is estimated as $g^*(\boldsymbol{\theta}_i|\boldsymbol{t}_i) \sim N\left(\boldsymbol{t}_i^{\boldsymbol{T}}\Gamma^*, \Sigma^*\right)$. In the case of PISA, the number of covariates is reduced to a number of principal components sufficient to retain 80% of the variance in the original full set of indicators or so that the number of retained components does not exceed 5% of the sample size within a country (OECD, 2017, p. 182). From this author's attempt to recreate the process with the UK's data, this results in the thousands of background indicators being compressed down into roughly 250 principal components.

Amongst the thousands of variables in the conditioning model that are compressed by the PCA, are indicators of the school that each student attends. These variables are of particular interest for this research study as they are the only part of the conditioning model that indicates the region of the UK in which each student attends school. Thus, along with all the other variables, information about whether a student attends school in England or in Northern Ireland will not be retained directly but will instead be condensed into the principal components.

Crucially, it should be noted that the PCA is conducted using weighted data and based upon a covariance rather than a correlation matrix. This means that retaining data from background indicators that are prevalent in the sample but, after weighting, are not estimated to be prevalent in the full population will be considered a lower priority. This distinction is important for thinking about the analysis of data from Northern Ireland. Students from Northern Ireland are deliberately oversampled so that whilst 2401 students from Northern Ireland participated in PISA 2015 (around 17% of the UK's sample), using the official weights in the PISA data, Northern Ireland's students are only estimated to comprise 3% of the UK's 15-year-old school population. Thus, after weighting, an indicator variable of attendance at a particular school in Northern Ireland will have an extremely low mean (and variance). This fact leads to a risk that the PCA will assign little priority to retaining information about attendance at schools within Northern Ireland. As such, the conditioning model may end up including information about individual schools within England, but not account for the effect of individual schools in Northern Ireland. This could, in theory, lead to an underestimation of the variation in student abilities in Northern Ireland as multiple imputation based upon the conditioning model essentially assumes that such effects have somewhat been accounted for when, in fact, they have not.

The remainder of this paper comprises a simulation study to illustrate this possible effect and empirical analysis showing the difference it makes to Northern Ireland's results if the principal components step in creating plausible values is avoided. Note that, prior to PISA 2015, school IDs were explicitly included in the conditioning model without being preprocessed using PCA (OECD, 2014, p. 157). Thus, the findings in this paper do not relate to the PISA datasets from 2012 or earlier.

3 Simulation Study

A simulation study was used to illustrate the possible effect of using principal components upon background variables including school IDs as a precursor to fitting the conditioning model. For simplicity, the simulation is based on measurement from a unidimensional test. To begin with, the abilities of students (θ_$_{\text{ijk}}$) were simulated according to the following set of equations:

$$\theta_{ijk} \sim N\big(\mu_{jk} + y_{ijk}, 0.65\big) \tag{3}$$

$$\mu_{jk} \sim N(\beta_k, 0.25)\ \{\beta_0 = 0.0, \beta_1 = 0.2\} \tag{4}$$

$$y_{ijk} = N(0, 0.10) \tag{5}$$

In the above formulae, θ_{ijk} represents the ability of the ith student in the jth school within the kth region, μ_{jk} is the effect of the jth school in the kth region on ability, and β_k is the mean school effect within the kth region. Only two regions are included in the simulation so that k can take the values 0 or 1. In region 0, the mean school effect is set to be zero, and in region 1, the mean school effect is set to be 0.2. y_{ijk} is a continuous background variable that explains a small proportion of the variation in the abilities of students within schools. Abilities were simulated for 25 students in each of 300 schools in region 0 and 100 schools in region 1. This means that the simulated data set contained 7500 students in region 0 and 2500 in region 1. Throughout the analysis, data within region 0 was given 10 times as much weight as in region 1.

The way this simulation study was set up was chosen to broadly reflect the combined PISA data set in England and Northern Ireland. For example, the real PISA 2015 data set contained 2400 students from Northern Ireland and just over 5000 from England (plus just over 3000 from Wales which, in practice, are also analyzed concurrently). Similarly, the combined data was drawn from 95 schools in Northern Ireland and over 200 in England (plus more than 100 in Wales). Again, the mean student weight from PISA 2015 students in England is 12 times that from students in Northern Ireland. Equations 3–5 also imply that the overall variance in ability will be 1 and that the intraschool correlation in ability will be roughly 0.25. This is very close to the estimated value of the intraschool correlation in England in the real PISA data.

The difference between regions (0.2) was chosen to be reasonably large in order to make any bias in the estimated difference in means visible against the likely standard errors. The background variable y_{ijk} was included as it would be highly unusual to apply PCA to a data set with just a single categorical variable (school ID). The variance of y_{ijk} was chosen to yield a correlation between y_{ijk} and ability of just above 0.3; a similar value to, for example, the correlation between socio-economic status and reading performance in PISA (OECD, 2013, p. 175). Only a single background variable was included in the simulation study in order to keep this illustrative example as simple as possible.

Using the simulated abilities, scores on 30 dichotomous items were simulated for each student using a standard Rasch model. Item difficulties were set up to be equally spaced between -1 and $+1$. The number of items was chosen to be roughly the same as the number taken by students within each PISA domain (if they take any within the domain at all). The reliability of the simulated test scores was found to be roughly 0.85 (using coefficient alpha) which is similar to the level of reliability reported for PISA test scores (OECD, 2017, p. 231).

Now using data from the item scores, as well as (optionally) the background information (school IDs and the values from y_{ijk}), the aim of analysis was to compare the distribution of estimated abilities between regions. Five methods of analysis were used:

1. Direct latent variable regression to estimate the relationship between region and ability. For the purposes of the charts this method is denoted "LV". Specifically the likelihood of each student's set of item responses ($\boldsymbol{x_{ijk}}$) was defined by the equation below where the item difficulty parameters (d_m) were fixed at their known values.

$$f\left(\boldsymbol{x_{ijk}}|\theta_{ijk}\right) = \prod_m f_m\left(x_{mijk}|\theta_{ijk}\right) = \prod_m \frac{\exp\left(\theta_{ijk} - d_m\right)^{x_{mijk}}}{1 + \exp\left(\theta_{ijk} - d_m\right)} \tag{6}$$

The likelihood values for each student across a range of abilities were estimated in R using the function "IRT.likelihood" from the package TAM (Robitzsch et al., 2017). Once this was done, it was possible to directly estimate a latent variable regression using the function "tam.latreg" using the formula given below.

$$\theta_{ijk} = \widehat{\beta_k} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N\left(0, \hat{\sigma}^2\right) \tag{7}$$

The $\widehat{\beta_k}$ coefficients are used to provide a direct estimate of the mean difference between regions.

2. Use of IRT to produce a single EAP estimate of each student's ability and compare these EAP scores between regions. This method is denoted "EAP". The EAP estimates were derived by combining the likelihood function defined in Eq. (6) with a normally distributed prior for ability. The mean and standard deviation of this prior were estimated using the function "tam.mml".
3. Analysis of plausible values produced using IRT combined with a conditioning model relating ability to school IDs and to the background variable. This method is denoted "PV-FULL". For this method, the likelihood was defined as in Eq. (6) but a latent regression model was used to account for the effect of school IDs (not just regions) and the background variable using the formula below.

$$\theta_{ijk} = \widehat{\mu_{jk}} + \hat{\alpha} y_{ijk} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N\left(0, \hat{\sigma}^2\right) \tag{8}$$

The likelihood and the estimated regression parameters were jointly used to produce plausible values using the function "tam.pv". Five plausible values were created for each student.

4. Analysis of plausible values produced using IRT combined with conditioning model relating ability to principal components of (indicators of) school IDs and to the background variable designed so that 80% of the original variance is retained. This method is denoted "PV-PCA". For this method, the set of school IDs were contrast coded to create 400 indicator variables. The variable y_{ijk} was also contrast coded into quintiles to create five further indicator variables. Contrast coding was used as it reflects the way most variables are handled in the conditioning model used in PISA (OECD, 2017, p. 182). The covariance matrix for the resulting 405 variables was calculated using weighted data and sufficient principal components (which we will denote $\boldsymbol{t_{ijk}}$) were extracted to explain 80% of the variance. A latent regression model was estimated using the equation below.

$$\theta_{ijk} = \widehat{\mu_0} + \boldsymbol{t}_{\boldsymbol{ijk}}^{\boldsymbol{T}}\widehat{\Gamma} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N\left(0, \hat{\sigma}^2\right) \tag{9}$$

The likelihood and the estimated regression parameters were jointly used to produce plausible values using the function "tam.pv". Five plausible values were created for each student.

5. Analysis of plausible values produced from IRT without including any covariates at all in the conditioning model. This method is denoted "PV-NULL". For this method, five plausible values were generated for each student based upon combining the likelihood function with the (very) simple latent regression equation defined below.

$$\theta_{ijk} = \widehat{\mu_0} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N\left(0, \hat{\sigma}^2\right) \tag{10}$$

For the purposes of this analysis, all IRT models were fitted using marginal maximum likelihood (MML) estimation. The simulated data was created 1000 times including the simulation of students' abilities (Eqs. 3–5) and item scores. All five forms of analysis were applied to each of these 1000 simulated data sets.

The results of the analysis are shown in Figs. 3 and 4. Figure 3 shows the distribution of the estimated mean differences between regions from the five methods across the 1000 simulations. The dotted line represents the true value of this difference used within simulations of 0.2. As can be seen, both direct latent variable modelling and the use of plausible values derived using a full conditioning model produced approximately unbiased estimates overall. In contrast, each of the other three methods yielded estimated differences in means that are biased downwards. However, it is interesting to note that, in this case the approach based upon EAP ability estimates is no worse than using plausible values either from the reduced model (using principal components only) or without including any covariates in the conditioning model at all. These results are not unexpected. Indeed, the results here

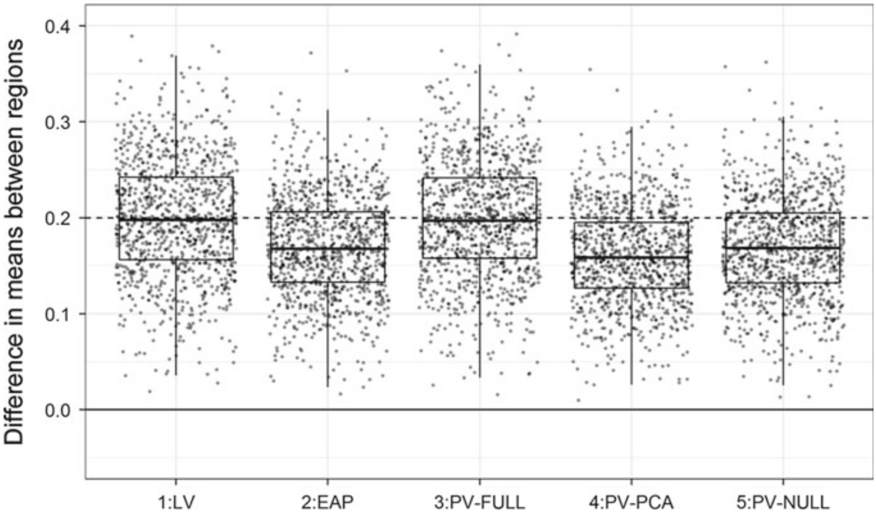


Fig. 3 Distribution of estimated mean differences in ability between regions using five different methods across 1000 simulations

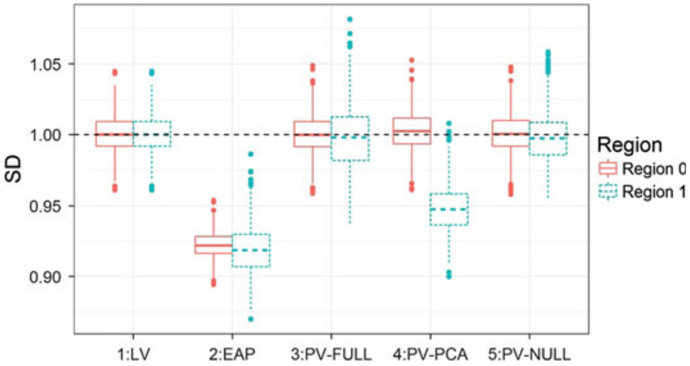


Fig. 4 Distribution of estimated standard deviations of ability in each region using five different methods across 1000 simulations

illustrate why it is important that a conditioning model is used in the way it is in the PISA studies. Failing to use a conditioning model leads to underestimating the difference between regions. Also, as suspected, due to the weighting of the data, the PCA fails to capture attendance at schools in region 1. As such, this reduced conditioning model does no better job at producing unbiased estimates than omitting all covariates from the conditioning model.

Figure 4 shows the distribution of estimated standard deviations within each region from each method. All methods, with the exception of using EAP, provide approximately unbiased estimates of the standard deviation of abilities within region 0. Three of the methods (direct latent regression, plausible values based on a full model, and plausible values with no covariates in the conditioning model) also provide unbiased estimates within region 1. Note that, in this analysis, the direct latent regression model only estimates one variance parameter and so always estimates the same variance within each region—a fact that is advantageous in this simulation but may not be in others.

Of most interest for our research paper, is the fact that, as expected, the method where PCA is used a precursor to the conditioning model badly underestimates the standard deviation of ability within region 1. This is because, due to the weighting of the data, the PCA fails to capture information about which schools students attend within region 1. As such, school effects are included in imputation to some extent within region 0 but hardly included within region 1. This results in an additional spread of plausible values between schools in region 0 but not in region 1.

Note that, although the EAP method clearly underestimates the true standard deviation in ability within each region, at least it is consistent. In particular, it correctly identifies that there is no substantive difference in the spread of abilities between the two regions. This is important, as it is the *comparison* between countries, regions and over time that is likely to be of substantive interest to policy makers. As long as whatever method we use treats these different entities consistently, we may not care if it is biased in a strictly psychometric sense. In this example, the substantive question is whether the spread of abilities differs between regions. The actual standard deviation of abilities is on a fairly arbitrary IRT scale, and it does not matter whether the actual values match those used to create the simulation.

4 The Effect of the Conditioning Model on Comparisons of England and Northern Ireland

Finally, we examined the actual empirical item-level data from PISA 2015 to see if different approaches to creating ability estimates might change the conclusions around the spread of ability in Northern Ireland as compared to England. In particular, it was of interest to explore whether generating reading plausible values without using PCA as a precursor to the conditioning model might yield a different set of results.

The analysis compared estimates of the mean, 10th percentile and 90th percentile of ability using the official PISA plausible values to four further ways of creating ability estimates:

1. An attempt to, as far as possible, recreate the methodology described in OECD (2017). This included contrast coding of all background variables, PCA and multidimensional IRT using a conditioning model. The major difference was that analysis was completed in R using the package TAM (Robitzsch et al., 2017). The purpose of including these ability estimates was to ensure that other reported differences were indeed due to the approach to the conditioning model rather than due to changes in software.

Table 1 Estimated means and percentiles of PISA reading abilities in England and Northern Ireland

Method	Mean		10th percentile		90th percentile	
	Eng.	Nor. Ire.	Eng.	Nor. Ire.	Eng.	Nor. Ire.
Official PVs	499.6	497.0	371.2	385.0	625.2	604.8
PVs—recreated methodology	500.9	498.9	377.7	387.0	621.6	609.4
PVs—no conditioning, separate countries	499.5	501.8	376.8	385.0	618.7	615.4
PVs—conditioning on school IDs only	500.5	503.9	375.5	381.9	620.3	625.6
EAP ability estimates	500.5	504.7	384.2	388.4	612.1	614.6

2. Plausible values derived without any conditioning model, but with the central IRT model fitted separately for England and Northern Ireland. This is the most direct approach to estimating separate ability distributions in each country.
3. Plausible values based on a conditioning model without preprocessing using PCA. Note that avoiding using PCA means that we cannot include all of the possible background variables in the conditioning model and instead limit ourselves to including school IDs only.
4. EAP estimates of ability based upon unidimensional IRT and no conditioning model. These estimates are the closest we get to simply using a single test score for each student. Note that each student must have answered at least one item in a domain in order to be included in the estimate of the ability distribution.

The results of analysis are shown in Table 1. Although multidimensional IRT was used in the generation of all plausible values, for brevity, only the results for reading are shown. To begin with, we note that, for England, all methods produced very similar estimates of mean ability. It can also be seen that the attempt to reproduce the official methodology resulted in statistics fairly close to the official ones both for the mean and at the two percentiles of interest.

All three of the alternative methods resulted in slightly higher mean abilities in Northern Ireland for reading. This may be because each of these methods avoided the inference that because, as is a fact, Northern Ireland performed worse in science than England, it must also perform worse in reading and math.

As expected, both alternative methods of producing plausible values led to larger differences between the 10th and 90th percentile in Northern Ireland. This may be because either separate estimation in each region, or using the school IDs as a direct conditioning variable without preprocessing using PCA, allows a greater degree of between-school variation to be retained in the imputations. In particular, the abilities of students in Northern Ireland at the 90th percentile are substantially higher using either of these alternative approaches to producing plausible values. This suggests that the current process based on a combination of PCA and a conditioning model may have led to biased estimates of these percentiles in Northern Ireland.

5 Discussion

This paper has explored the impact of the use of principal components analysis as a pre-cursor to the production of plausible values in PISA. It has shown that, in very particular circumstances, this can potentially lead to bias in the estimated distributions of ability. The findings may have particular substantive importance for secondary analysis of subpopulations that are deliberately oversampled in the PISA studies.

These findings contrast with the generally positive tone regarding plausible values in the psychometric literature (e.g., Braun & Von Davier, 2017, Marsman et al., 2016). The results show how this well-designed methodology, that provides excellent performance in simulation studies, may fail to retain all of its desired properties when faced with real, large-scale data sets and the associated challenges of these, such as the need for data reduction.

Having said the above, the motivation for using a combination of PCA and a conditioning model within the PISA studies is correct. The use of a conditioning model has theoretical benefits in ensuring the accuracy of ability estimates but cannot be completed without PCA given the vast amount of background information collected on each student. As such, this paper is not intended as a criticism of the core methodology that has been adopted in PISA. However, it is clear that in some cases, minor methodological details can have an influence upon substantive results. With this in mind, it is important that analysts understand these methodologies and the alternative approaches that are available to them. It is notable that while a great many software tools have been produced to facilitate the analysis of plausible values, far less has been done to help analysts create them in their own contexts. This becomes a particularly important issue if PISA data is matched to other sources of information that were not included in the main conditioning model.

It may be that expecting data analysts around the world to become comfortable with creating their own sets of plausible values is asking too much. With this in mind, a sensible recommendation for analysts would be that it is worth checking surprising results from the PISA studies against simpler measures of ability such as EAP ability estimates.

References

Braun, H., & von Davier, M. (2017). The use of test scores from large-scale assessment surveys: Psychometric and statistical considerations. *Large-scale Assessments in Education, 5*(1), 17.

Chalmers, R. P. (2012). Mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1–29.

Jerrim, J., & Shure, N. (2016a). *Achievement of 15-Year-Olds in England: PISA 2015 National Report (DFE-RR630)*. London, England: Department for Education.

Jerrim, J., & Shure, N. (2016b). *Achievement of 15-Year-Olds in Northern Ireland: PISA 2015 National Report*. London, England: UCL.

Laukaityte, I., & Wiberg, M. (2017). Using plausible values in secondary analysis in large-scale assessments. *Communication in Statistics—Theory and Methods, 46*(22), 11341–11357.

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2016). What can we learn from Plausible Values? *Psychometrika, 81*(2), 274–289.

OECD. (2013). *PISA 2012 results: Excellence through equity: Giving every student the chance to succeed* (Vol. II). Paris: OECD.

OECD. (2014). *PISA 2012 technical report*. Paris: OECD.

OECD. (2017). *PISA 2015 technical report*. Paris: OECD.

Robitzsch, A., Kiefer, T., & Wu, M. (2017). TAM: Test analysis modules (Version 2.8–21) [Computer software]. Retrieved from http://CRAN.R-project.org/package=TAM.

Von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. *IERI monograph series: Vol. 2*, (pp. 9–36).

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2–3), 114–128.

Adopting the Multi-process Approach to Detect Differential Item Functioning in Likert Scales

Kuan-Yu Jin, Yi-Jhen Wu and Hui-Fang Chen

Abstract The current study compared the performance of the logistic regression (LR) and the odds ratio (OR) approaches in differential item functioning (DIF) detection in which the three processes of an IRTree model were considered in a five-point response scale. Three sets of binary pseudo items (BPI) were generated to indicate an intention of endorsing the midpoint response, a positive/negative attitude toward an item, and a tendency of using extreme category, respectively. Missing values inevitably appeared in the last two sets of BPI. We manipulated the DIF patterns, the percentages of DIF items, and the purification procedure (with/without). The results suggested that (1) both the LR and OR performed well in detecting DIF when BPI did not include missing values; (2) the OR method generally outperformed the LR method when BPI included missing values; (3) the OR method performed fairly well without a purification procedure, but the purification procedure improved the performance of the LR approach, especially when the number of DIF was large.

Keywords IRTree · Differential item functioning · Logistic regression · Odds ratio · Purification · Missing data

1 Introduction

Item response tree (IRTree) models (Böckenholt, 2012) have become popular recently because of two appealing features. First, IRTree models apply an intuitive approach by visualizing underlying response processes through tree-like structures. Second,

K.-Y. Jin (✉)
Faculty of Education, The University of Hong Kong, Pokfulam, Hong Kong, China
e-mail: kyjin@hku.hk

Y.-J. Wu
Bamberg Graduate School of Social Sciences (BAGSS), Otto-Friedrich-Universität Bamberg, Feldkirchenstr. 21, 96052 Bamberg, Germany

H.-F. Chen
Department of Social and Behavioural Sciences, City University of Hong Kong, Tat Chee Ave, Kowloon, Hong Kong, China

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_27

IRTree models simultaneously model both the content-related trait and different types of response processes, and make it possible to investigate more than one type of response process at a time (Plieninger & Heck, 2018; Zettler, Lang, Hülsheger, & Hilbig, 2016).

To the best of our knowledge, researchers have not yet addressed differential item function (DIF) issues under the IRTree framework. By recognizing the popularity of non-IRT approaches to identifying DIF in missing data (Emenogu, Falenchuck, & Childs, 2010; Jin, Chen, & Wang, 2018), this study aimed to compare the performance of the logistic regression (LR; Rogers & Swaminathan, 1993) and the odds ratio (OR; Jin et al., 2018) approaches to identify DIF in the IRTree models. Because a five-point response scale is commonly used in self-reported instruments, we focused on this format throughout the study.

1.1 IRTree Model

We adopted the three-process IRTree model (Böckenholt, 2012) as an example (Fig. 1). The observed response on an item is derived from multiple processes that give a rise to the selection of one of response categories. Each process is assumed to measure a latent variable and modeled by a binary pseudo item (BPI). A BPI has two branches attached with transition probabilities: Pr and $1 - \text{Pr}$. In Process I, referred to indifferences, individuals decide if they have an opinion to an item or want to express their attitude. If respondents do not have a clear opinion or refuse to indicate their attitudes, they would endorse the midpoint (i.e. *neutral*) and stop at Process I. A transition probability of Process I is denoted as Pr_1. Those who do not endorse the midpoint would keep proceeding to Process II and Process III. At Process II, indicating direction, individuals decide to agree or disagree with an item. If participants hold a positive attitude, they would endorse the item and a transition probability is Pr_2. If individuals have a negative attitude, they would choose disagreement and the transition probability of Process II is $1 - \text{Pr}_2$. Lastly, in Process III, individuals decide the intensity of attitude (i.e. *strongly agree* vs. *agree or strongly disagree* vs *disagree*) and then endorse the extreme or less extreme category. Participants with a strong attitude would choose the extreme category, and a transition probability is Pr_3; whereas the ones with a weak attitude would endorse the less extreme response, and a transition probability is $1 - \text{Pr}_3$. Given the assumption of local independence, the probability of a final observable response is obtained by multiplying three transition probabilities.

The left half of Table 1 depicts how original responses are decomposed into three BPI. It is noticeable that missing values inevitably occur in Process II and Process III. For example, an individual endorses the category 2, and his or her response process will not proceed Process II and Process III. Consequently, no information is available on BPI of Process II and Process III, respectively.

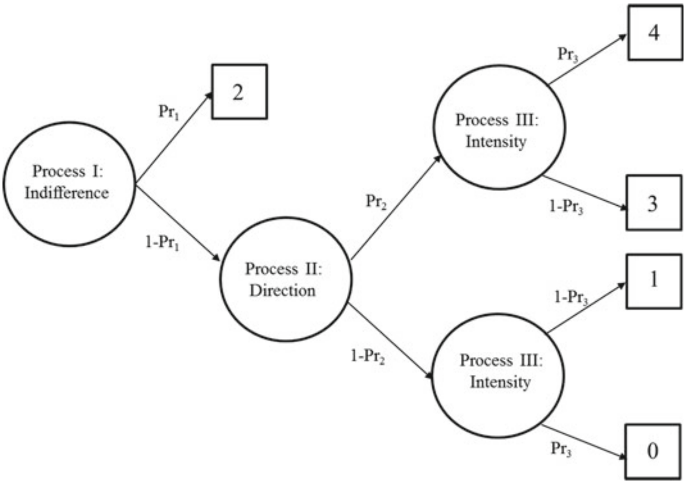


Fig. 1 The tree-structure for an item with a five-point response scale (0 to 4; Strongly disagreement to Strongly agreement)

Table 1 Pseudo-items of the five-point response scale

	BPI coding			
Response	I	II	III	Probability
0	0	0	1	$\frac{1}{1+\exp\left(\theta_i^{\mathrm{I}}-\beta_j^{\mathrm{I}}+\gamma_j^{\mathrm{I}}\right)} \times \frac{1}{1+\exp\left(\theta_i^{\mathrm{II}}-\beta_j^{\mathrm{II}}+\gamma_j^{\mathrm{II}}\right)} \times \frac{\exp\left(\theta_i^{\mathrm{III}}-\beta_j^{\mathrm{III}}+\gamma_j^{\mathrm{III}}\right)}{1+\exp\left(\theta_i^{\mathrm{III}}-\beta_j^{\mathrm{III}}+\gamma_j^{\mathrm{III}}\right)}$
1	0	0	0	$\frac{1}{1+\exp\left(\theta_i^{\mathrm{I}}-\beta_j^{\mathrm{I}}+\gamma_j^{\mathrm{I}}\right)} \times \frac{1}{1+\exp\left(\theta_i^{\mathrm{II}}-\beta_j^{\mathrm{II}}+\gamma_j^{\mathrm{II}}\right)} \times \frac{1}{1+\exp\left(\theta_i^{\mathrm{III}}-\beta_j^{\mathrm{III}}+\gamma_j^{\mathrm{III}}\right)}$
2	1	*	*	$\frac{\exp\left(\theta_i^{\mathrm{I}}-\beta_j^{\mathrm{I}}+\gamma_j^{\mathrm{I}}\right)}{1+\exp\left(\theta_i^{\mathrm{I}}-\beta_j^{\mathrm{I}}+\gamma_j^{\mathrm{I}}\right)}$
3	0	1	0	$\frac{1}{1+\exp\left(\theta_i^{\mathrm{I}}-\beta_j^{\mathrm{I}}+\gamma_j^{\mathrm{I}}\right)} \times \frac{\exp\left(\theta_i^{\mathrm{II}}-\beta_j^{\mathrm{II}}+\gamma_j^{\mathrm{II}}\right)}{1+\exp\left(\theta_i^{\mathrm{II}}-\beta_j^{\mathrm{II}}+\gamma_j^{\mathrm{II}}\right)} \times \frac{1}{1+\exp\left(\theta_i^{\mathrm{III}}-\beta_j^{\mathrm{III}}+\gamma_j^{\mathrm{III}}\right)}$
4	0	1	1	$\frac{1}{1+\exp\left(\theta_i^{\mathrm{I}}-\beta_j^{\mathrm{I}}+\gamma_j^{\mathrm{I}}\right)} \times \frac{\exp\left(\theta_i^{\mathrm{II}}-\beta_j^{\mathrm{II}}+\gamma_j^{\mathrm{II}}\right)}{1+\exp\left(\theta_i^{\mathrm{II}}-\beta_j^{\mathrm{II}}+\gamma_j^{\mathrm{II}}\right)} \times \frac{\exp\left(\theta_i^{\mathrm{III}}-\beta_j^{\mathrm{III}}+\gamma_j^{\mathrm{III}}\right)}{1+\exp\left(\theta_i^{\mathrm{III}}-\beta_j^{\mathrm{III}}+\gamma_j^{\mathrm{III}}\right)}$

Note An asterisk represents missing values. θ_i^{I}, θ_i^{II}, and θ_i^{III} are the three latent traits of individual *i* on the BPI. β_i^{I}, β_i^{II}, and β_i^{III} are the difficulties of BPI of virtual item *i*; and γ_j^{II}, γ_j^{II}, and γ_j^{III} are the DIF parameters, which are the item difficulty differences between the focal group and the reference group on BPI

1.2 DIF in the IRTree Model

DIF analyses methods are used to investigate if an item is measuring different proficiencies for respondents of separate groups. In a traditional practice, researchers are interested in whether respondents from a reference group get a higher (or lower) item score than others from a focal group, given that they have the same proficiency level. However, it is commonly relatively difficult to understand why a DIF item is unfair to one group. In IRTree models nuisance factors correlated with the group membership may influence the response processes of endorsing the midpoint response, exhibiting the direction of attitude toward the statement, and choosing the extreme or less extreme category. Fortunately, the IRTree modelling provides more detailed information for possible sources of DIF for those response process. One can examine if multiple BPI, which present different response processes, are invariant across groups. A simpler case is that only one BPI is suspected of functioning differentially in one response process with respect to the group membership (e.g., gender). On the other hand, the group membership might be a crucial factor inducing DIF in more than one response processes. Either DIF patterns are of concern in this study.

Theoretically, DIF denotes that an additional dimension is not considered, and different groups of respondents have unequal means on the additional dimension (Walker & Sahin, 2017). The right half of Table 1 shows that γ_j^{I}, γ_j^{II}, and γ_j^{III} are the DIF parameters indicating group differences in the latent trait distributions for the additional dimension(s) on BPI; θ_i^{I}, θ_i^{II}, and θ_i^{III} are the latent traits of individual *i* measured by the corresponding BPI; and β_i^{I}, β_i^{II}, and β_i^{III} are the difficulties of BPI of an item *j* for the reference group. In this case, eight ($= 2^3$) possible combinations of DIF may occur (Table 2). When $\gamma_j^{\mathrm{I}} = \gamma_j^{\mathrm{II}} = \gamma_j^{\mathrm{III}} = 0$, suggesting that none of BPI is flagged as DIF (i.e., Pattern 1), the detected item is deemed DIF-free. An item is classified as *distinct DIF* when one of the γ-parameters of item *i* is not zero (i.e., Patterns 2–4); and an item is classified as c*oincided DIF* when more than one BPI is variant between groups (i.e., Patterns 5–8).

Table 2 Overview of DIF conditions in the IRTree framework

		BPI coding		
Pattern	Situation	I	II	III
1	DIF-free	0	0	0
2	Distinct DIF	1	0	0
3	Distinct DIF	0	1	0
4	Distinct DIF	0	0	1
5	Coincided DIF	1	1	0
6	Coincided DIF	1	0	1
7	Coincided DIF	0	1	1
8	Coincided DIF	1	1	1

Note 1 represents as the occurrence of DIF and 0 otherwise

1.3 DIF Approaches

Two schools of approaches have been proposed to detect DIF, including the IRT and non-IRT approach. Although both approaches perform well, non-IRT approaches are easily implemented in empirical studies (Lei, Chen, & Yu, 2006; Jin et al., 2018) and they are the focus of the present study.

Studies have reported that the performance of non-IRT approaches might be influenced by missing data (Emenogu et al., 2010). Several methods have been proposed to enhance their performance in missing data, such as listwise deletion and imputation. However, the two methods may not work well in missing data in DIF assessments (Jin et al., 2018), and we did not delete or impute data in Process II or III.

1.3.1 The Logistic Regression (LR) Approach

In the LR approach, coefficients for test scores (X), the group variable ($G = 0$ for the reference group and $G = 1$ for the focal group) and their interaction (XG) are tested for uniform and non-uniform DIF:

$$\log\left(\frac{P_{i1}}{P_{i0}}\right) = \mathrm{b}_{0i} + \mathrm{b}_{1j}X + \mathrm{b}_{2j}G + \mathrm{b}_{3j}XG, \tag{1}$$

where P_{j1} and P_{j0} are the probabilities of success and failure on item j, respectively; b_{0j} is an intercept for the item j; b_{1j} is the effect of test scores on item j; b_{2j} is used to capture a grouping impact as uniform DIF; and a significant b_{3j} signifies non-uniform DIF. The present study only focused on uniform DIF.

1.3.2 The Odds Ratio (OR) Approach

In the OR method, $\hat{\lambda}_j$ represents the logarithm of the OR of success over failure on item i for the reference group and the focal group as following:

$$\hat{\lambda}_j = \log\left(\frac{\mathrm{n}_{\mathrm{R}1j}/\mathrm{n}_{\mathrm{R}0j}}{\mathrm{n}_{\mathrm{F}1j}/\mathrm{n}_{\mathrm{F}0j}}\right), \tag{2}$$

where $\mathrm{n}_{\mathrm{R}1j}$ and $\mathrm{n}_{\mathrm{R}0j}$ are the number of individuals answering item j correctly and incorrectly in the reference group; $\mathrm{n}_{\mathrm{F}1j}$ and $\mathrm{n}_{\mathrm{F}0j}$ are the number of individuals answering item j correctly and incorrectly in the focal group. $\hat{\lambda}_j$ follows a normal distribution asymptotically, with a mean of λ and standard deviation of

$$\sigma\left(\hat{\lambda}_j\right) = \sqrt{\mathrm{n}_{\mathrm{R}1j}^{-1} + \mathrm{n}_{\mathrm{R}0j}^{-1} + \mathrm{n}_{\mathrm{F}1j}^{-1} + \mathrm{n}_{\mathrm{F}0j}^{-1}}. \tag{3}$$

Given a nominal level of α, the confidence interval of $\hat{\lambda}_j$ (given a nominal level of α) is used to examine if an item *j* has DIF. If $\hat{\lambda}_j \pm z_{\alpha/2} \times \sigma(\hat{\lambda}_j)$ does not contain λ, item *j* is deemed as having DIF. Jin et al. (2018) recommended the sample median of $\hat{\lambda}_j$ as the estimator of λ. Consequently, an extreme large $\hat{\lambda}_j$ indicates that item *i* favors the reference group, whereas an extreme small $\hat{\lambda}_j$ indicates item *j* favors the focal group. Note that the value of $\sigma(\hat{\lambda}_j)$ is dependent of the size of non-missing data. The more the missing data (i.e. Process II or III) occur, the larger the confidence interval of $\hat{\lambda}_j$ would be. This feature helps the OR method be less influenced by missing data.

2 Simulation Study

2.1 Methods

A total of 1000 respondents (500 in the focal group and 500 in the reference group) were simulated to answer 30 five-point response scale items. Four factors were manipulated: (1) the percentage of DIF items: 10%, 20%, and 30%, (2) the pattern of DIF on BPI: distinct or coincided, (3) DIF direction: bilateral or unilateral, (4) purification procedure: with or without. In the bilateral condition, two thirds of DIF items favored the reference group, and one third of DIF items favored the focal group. In the unilateral condition, all DIF items favored the reference group. The purification procedure is an iterative procedure to identify DIF items until the identified DIF items at two successive iterations are identical (Wang & Su, 2004). A total of 100 replications were conducted under each condition.

The three traits (θ_i^{I}, θ_i^{II}, and θ_i^{III}) for the reference and focal groups were generated from an identical multivariate normal distribution, where $\boldsymbol{\mu}' = [0, 0, 0]$ and $\Sigma = \begin{bmatrix} 1 & -0.2 & -0.2 \\ -0.2 & 1 & 0.2 \\ -0.2 & 0.2 & 1 \end{bmatrix}$, in which θ_i^{I} negatively correlated with θ_i^{II} and θ_i^{III} and θ_i^{I} and θ_i^{II} were positively correlated. Item responses were generated in accordance with Table 1. Item difficulties followed a uniform distribution between 1.5 and -1.5. The DIF sizes were set as 0.25. The performance of the LR and OR methods were evaluated via the false positive rate (FPR) and the true positive rate (TPR). FPR is that a DIF-free item was mistakenly identified as a DIF item, and TPR is that a DIF item was correctly identified as a DIF item. The LR method was implemented with the *difLogistic* function from the *difR* package (Magis, Béland, Tuerlinckx, & De Boeck, 2010) in R version 3.2.5 (R Core Team, 2016). The OR method was implemented by writing an R script, which is available from the first author upon request.

2.2 Results

2.2.1 The Unilateral DIF Conditions

FPR substantially inflated in the LR as the number of DIF items increased, especially when 30% DIF occurred in the coincided DIF conditions (Fig. 2a, b). Although the purification procedure helped reduce FPR, the corrected FPR were above the nominal level of .05. The purification procedure worked well to reduce FPR in Process I because the sum score of BPI was not affected by missing data. The OR yielded satisfactory FPR across conditions, although FPR were slightly inflated when there were 30% DIF items in the coincided and distinct DIF conditions (Fig. 2c, d). Likewise, the OR yielded lower FPR in Process I than in Processes II and III. Furthermore, the effectiveness of a purification procedure was only noticeable in the conditions when 30% of items were coincided and distinct DIF items in Process I. In sum, the efficiency of the purification procedure was more salient in the LR.

The LR and OR yielded similar TPR in most conditions (Fig. 2e–h). As the number of DIF items increased, TPR decreased. Especially, both methods yielded the lowest TPR when there were 30% DIF items in the distinct DIF conditions. Artificial missingness in Processes II and III caused limited information for DIF detection and resulted in lower TPR than in Process I.

2.2.2 The Bilateral DIF Conditions

When DIF items did not uniformly favor one group, both the LR and OR yielded FPR close to .05 with and without the purification procedure (Fig. 3a–d). That is, to some extent the influence of DIF was canceled out in the bilateral DIF conditions and did not lead to inflated FPR. The OR yielded more conservative FPR than the LR. Although higher FPR appeared in the LR when 30% DIF items were generated, the magnitude of inflation was acceptable.

Overall, the TPR were higher in the coincided DIF conditions for the LR and the OR (Fig. 3e–h), which suggested that both methods were more efficient to detect the coincided DIF compared to the distinct DIF. The higher TPR were observed in Process I in both methods because the first set of BPI did not have missing values. In addition, the LR yielded higher TPR than the OR in most conditions.

3 Discussion

The current study filled the knowledge gap in IRTree models whereby we compared the performance of the LR and the OR methods in detecting DIF items in BPI. The findings showed that the OR yielded satisfactory performance in most conditions than the LR did. Specifically, the OR yielded well-controlled FPR in the uniliteral

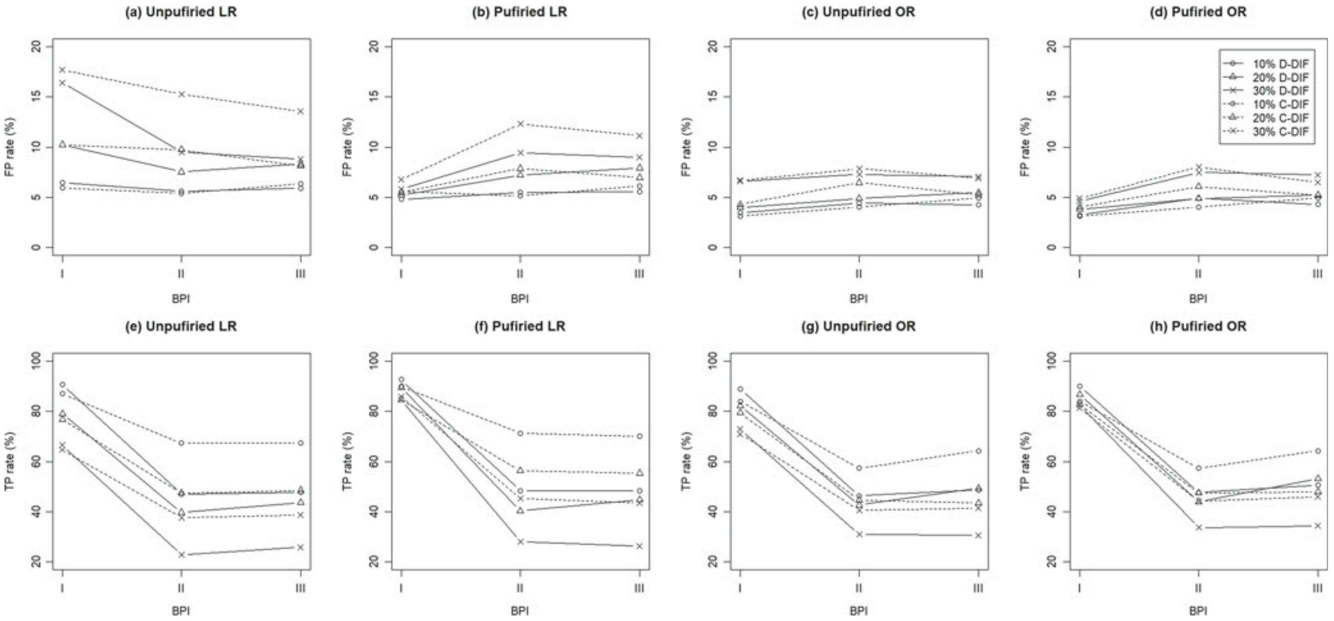


Fig. 2 FPR and TPR in the unilateral DIF conditions. *Note* D-DIF represents as the distinct DIF condition; C-DIF represents as the coincided DIF condition

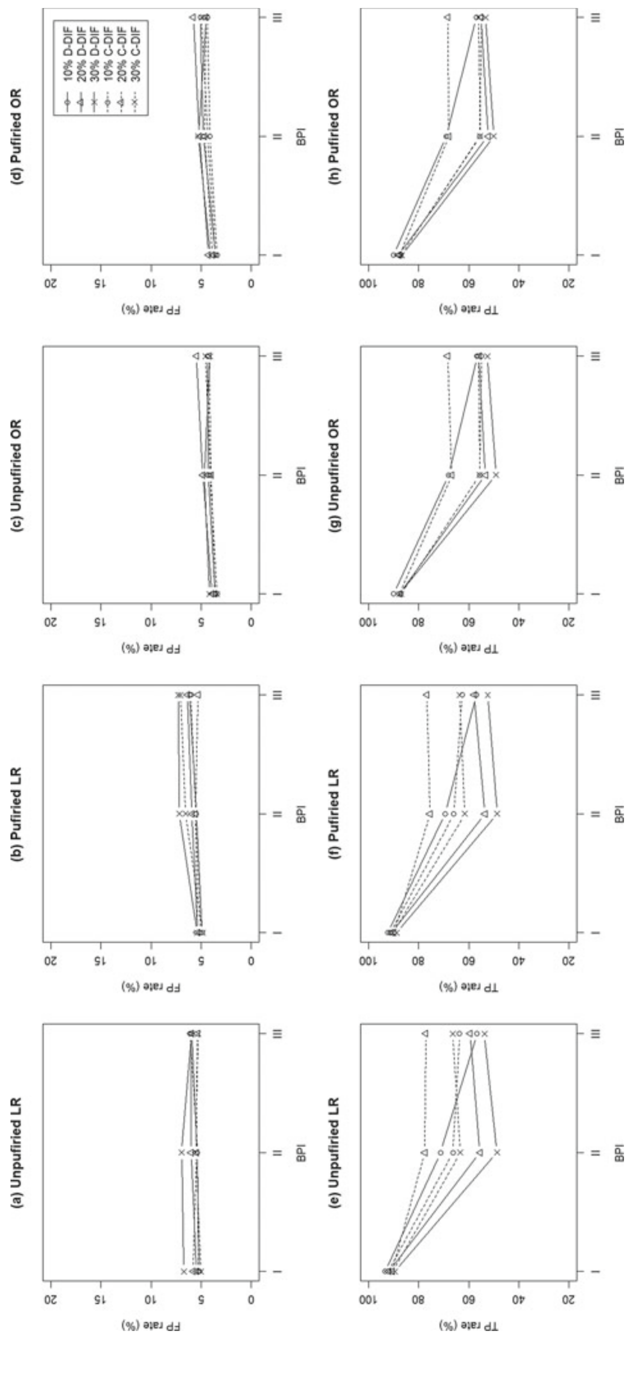


Fig. 3 FPR and TPR in the bilateral DIF conditions. *Note* D-DIF represents as the distinct DIF condition; C-DIF represents as the coincided DIF condition

conditions, but the LR did not. Although the LR could benefit from the purification procedure when a test contained 30% or above DIF items, the efficiency of a purification procedure was only salient in Process I. The low effectiveness of the purification procedure in Processes II and III is due to the joint occurrence of missingness and a large amount of DIF items. Regarding TPR, the OR and the LR had similar results in the uniliteral conditions; however, TPR was lowest in both methods when a test had 30% DIF.

The LR and OR yield satisfactory performance of FPR irrespective of the number of DIF items in the bilateral conditions. In the bilateral conditions where some DIF items favor the focal group and some DIF items favor the reference group, the effects of contamination on test scores were canceled out. In Process II or III, the sum of BPI was a valid indicator to match subjects in the LR, even though the number of DIF items is high (Wang & Su, 2004).

This study is not free of limitations. We only adopted the five-point response scale to evaluate the performance of the LR and OR methods in DIF detection. Future research should consider other situations, such like a four- or seven-point scale. Secondly, we only focused on the effect of uniform DIF on the BPI. To investigate nonuniform DIF, one can replace the item response function of BPI by the two-parameter logistic model, and then examine if the slope function differentially between groups. Lastly, the difference of latent proficiency between the two groups was not manipulated. Further studied should investigate whether or not the DIF detection result would be affected by unequal latent proficiency distributions.

References

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods, 17,* 665–678. https://doi.org/10.1037/a0028111.

Emenogu, B. C., Falenchuck, O., & Childs, R. A. (2010). The effect of missing data treatment on Mantel-Haenszel DIF detection. *The Alberta Journal of Educational Research, 56,* 459–469.

Jin, K.-Y., Chen, H.-F., & Wang, W.-C. (2018). Using odds ratios to detect differential item functioning. *Applied Psychological Measurement*. https://doi.org/10.1177/0146621618762738.

Lei, P., Chen, S., & Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement, 43,* 245–264. https://doi.org/10.1111/j.1745-3984.2006.00015.x.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42,* 847–862. https://doi.org/10.3758/brm.42.3.847.

Plieninger, H., & Heck, D. W. (2018). A new model for acquiescence at the interface of psychometrics and cognitive psychology. *Multivariate behavioral research, 53*, 633–654. https://doi.org/10.1080/00273171.2018.1469966.

Team, R. C. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17,* 105–116. https://doi.org/10.1177/014662169301700201.

Walker, C. M., & Sahin, S. G. (2017). Using a multidimensional IRT framework to better understand differential item functioning (DIF): A tale of three DIF detection procedures. *Educational and Psychological Measurement, 77,* 945–970. https://doi.org/10.1177/0013164416657137.

Wang, W.-C., & Su, Y.-H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education, 17,* 113–144. https://doi.org/10.1207/s15324818ame1702_2.

Zettler, I., Lang, J. W., Hülsheger, U. R., & Hilbig, B. E. (2016). Dissociating indifferent, directional, and extreme responding in personality data: Applying the three-process model to self-and observer reports. *Journal of Personality, 84,* 461–472. https://doi.org/10.1111/jopy.12172.

Detection of Differential Item Functioning via the Credible Intervals and Odds Ratios Methods

Ya-Hui Su and Henghsiu Tsai

Abstract Differential item functioning (DIF) analysis is an essential procedure for educational and psychological tests to identify items that exhibit varying degrees of DIF. DIF means that the assumption of measurement invariance is violated, and then test scores are incomparable for individuals of the same ability level from different groups, which substantially threatens test validity. In this paper, we investigated the credible intervals (CI) and odds ratios (OR) methods to detect uniform DIF within the framework of the Rasch model through a series of simulations. The results showed that the CI method performed better than the OR method to identify DIF items under the balanced DIF conditions. However, the CI method yielded inflated false positive rates under the unbalanced DIF conditions. The effectiveness of these two approaches was illustrated with an empirical example.

Keywords Credible interval · Odds ratio · DIF · Markov chain Monte Carlo · IRT

1 Introduction

Differential item functioning (DIF) analysis is an essential procedure for educational and psychological tests. DIF occurs when individuals from different groups (such as gender, ethnicity, country, or age) have different probabilities of endorsing or accurately answering a given item after controlling for overall test scores. It violates the assumption of measurement invariance and the test scores become incomparable for individuals of the same ability level from different groups, which substantially threatens test validity. DIF detection can examine how test scores are affected by

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

H. Tsai (✉)
Institute of Statistical Science, Academia Sinica, 128 Academia Road, Section 2, Nankang District, 11529 Taipei, Taiwan
e-mail: htsai@stat.sinica.edu.tw

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_28

external variables that are not related to the construct (Glas, 1998). Therefore, it is important to know if items are subject to DIF; that is, to know if the examinees are fairly measured.

Many approaches have been developed to perform DIF detection, and they can be classified into two categories (Magis, Béland, Tuerlinckx, & De Boeck, 2010): item response theory (IRT)-based and non-IRT-based approaches. The IRT-based approaches include the Lagrange multiplier test (Glas, 1998), the likelihood ratio test (Cohen, Kim, & Wollack, 1996), Lord's chi-square test (Lord, 1980), Raju's (1988) signed area method, etc. The IRT-based approaches require estimating item parameters for different groups. After comparing these item parameters of different groups, an item is identified as a DIF item if the item parameters are significantly different between groups. By contrast, the non-IRT-based approaches require neither specific forms for the IRT models nor large sample sizes (Narayanon & Swaminathan, 1996). The non-IRT-based approaches include the Mantel-Haenszel (MH; Holland & Thayer, 1988), logistic regression (LR; Rogers & Swaminathan, 1993), simultaneous item bias test (SIBTEST; Shealy & Stout, 1993) methods, etc.

Among the non-IRT-based approaches, the MH and LR methods perform well in flagging DIF items when the percentage of DIF items is not very high and there is no mean ability difference between groups (French & Maller, 2007; Narayanon & Swaminathan, 1996). A common feature of these two methods is that examinees from different groups are placed on a common metric based on the test scores, which are usually called matching variables. The use of the matching variables is critical for DIF detection (Kopf, Zeileis, & Strobl, 2015). If the matching variables are contaminated (i.e., consisting of DIF items), examinees with the same ability levels would not be matched well, and the subsequent DIF detection would be biased (Clauser, Mazor, & Hambleton, 1993). In practice, it is challenging to identify a set of DIF-free items as the matching variables for DIF detection, especially when the percentage of DIF items is high or when DIF magnitudes are large (Narayanon & Swaminathan, 1996; Rogers & Swaminathan, 1993).

To overcome this difficulty, the odds ratios (OR; Jin, Chen, & Wang, 2018) method was proposed to detect uniform DIF under various manipulated variables, such as different DIF pattern, impact, sample size, and with/without purification. Jin, Chen, and Wang (2018) found that the OR method without a purification procedure outperformed the MH and LR methods in controlling false positive rates (FPR) and obtaining high true positive rates (TPR) when tests contained high percentages of DIF items. Another recently developed IRT-based DIF detection method was the credible interval (CI) method proposed by Su, Chang, & Tsai (2018) to detect uniform and non-uniform DIF items under the Bayesian framework. Su et al (2018) found that the CI method performed well; however, only unbalanced DIF conditions and no impact (i.e., mean ability difference between the reference and focal groups was zero) were considered in their study.

A common feature of the CI and OR methods is that both methods perform DIF detection after constructing intervals. The OR method follows the frequentist approach, and constructs the confidence interval for the mean ability difference between the reference and focal groups. By contrast, the CI method follows the

Bayesian approach, and constructs the credible interval for the item difficulty difference between the reference and focal groups. See next section for more details. Because of the nature of the Bayesian framework, the CI method would need more time to perform DIF examination. Besides, the CI method assumes Rasch (1960) model is a correct model for the data. By contrast, the OR method does not require the specification of an IRT model; however, this method may not work when the number of examinees of any group is very small. Given the very different nature of these two newly developed methods, it is interesting to compare these two methods under the Rasch model. In this paper, we investigated the performance of the CI and OR methods to detect uniform DIF within the framework of the Rasch model through a series of simulation studies. The effectiveness of these two approaches was illustrated with an empirical example.

2 The CI and OR DIF Detection Methods

2.1 The CI Method

We first review the CI method proposed by Su, Chang, and Tsai (2018). Let Y_{pj} be the dichotomous response of examinee p on item j, where $p = 1, \ldots, P$, and $j = 1, \ldots, J$. Denote b_j and θ_p as the difficulty parameter for item j and the examinee ability parameter for examinee p, respectively. In the Rasch (1960) model, the probability of examinee p getting a correct response on item j is given by

$$\pi_{pj} = \mathrm{P}\big(Y_{pj} = 1|\theta_p, b_j\big) = \frac{1}{1 + e^{-\theta_p + b_j}}. \tag{1}$$

An item is flagged as DIF if the probability of answering the item correctly differs across different groups after controlling for the underlying ability levels. The CI method was proposed to perform DIF detection under a Bayesian estimation framework (Su et al., 2018). Consider the simplest case of two groups, hence, examinee p either belongs to the reference group ($g_p = 0$) or to the focal group ($g_p = 1$). Furthermore, each group has its own difficulty parameter. Then, Eq. (1) becomes

$$\pi_{pj} = \mathrm{P}\big(Y_{pj} = 1\big|g_p, \theta_p, b_j, d_j\big) = \begin{cases} \frac{1}{1+e^{-\theta_p+b_j}}, & g_p = 0, \\ \frac{1}{1+e^{-\theta_p+d_j}}, & g_p = 1, \end{cases} \tag{2}$$

where b_j and d_j are the difficulty parameters for the reference and the focal groups, respectively. Alternatively, the notations of Glas (1998) is adopted to rewrite Eq. (2) as

$$\pi_{pj} = \mathrm{P}\big(Y_{pj} = 1\big|g_p, \theta_p, b_j, \delta_j\big) = \begin{cases} \frac{1}{1+e^{-\theta_p+b_j}}, & g_p = 0, \\ \frac{1}{1+e^{-\theta_p+b_j+\delta_j}}, & g_p = 1. \end{cases} \tag{3}$$

Equation (3) implies that the responses of the focal group need an additional difficulty parameter δ_j. Therefore, the following hypothesis is considered:

$$H_0 : \delta_j = 0 \text{ versus } H_1 : \delta_j \neq 0.$$

Due to the complexity of the likelihood function, a Bayesian estimation method is used. Specifically, we follow closely the Bayesian approaches proposed by Chang, Tsai, and Hsu (2014), Chang, Tsai, Su, and Lin (2016), and Su et al. (2018). In particular, a two-layer hierarchical prior is assumed for the model parameters to reduce the impact of the prior settings on the posterior inference. For model identification, we follow Frederickx, Tuerlinckx, de Boeck, and Magis (2010)'s paper by assuming that the marginal distribution of θ_p is normal:

$$\theta_p \sim \begin{cases} N\left(0, \sigma_r^2\right), & g_p = 0, \\ N\left(\mu_f, \sigma_f^2\right), & g_p = 1. \end{cases}$$

For the first-layer prior settings for the parameters, we assume

$$\begin{aligned} b_j &\sim N\left(\mu_b, \sigma_b^2\right), \\ d_j &\sim N\left(\mu_d, \sigma_d^2\right). \end{aligned}$$

Given the first-layer prior, we assume the second-layer prior to be

$$\begin{aligned} \mu_f &\sim N\left(\mu_1, \sigma_1^2\right), \\ \mu_b &\sim N\left(\mu_2, \sigma_2^2\right), \\ \mu_d &\sim N\left(\mu_3, \sigma_3^2\right), \\ \sigma_r^2 &\sim \text{Inv-Gamma}(\alpha_1, \beta_1), \\ \sigma_f^2 &\sim \text{Inv-Gamma}(\alpha_2, \beta_2), \\ \sigma_b^2 &\sim \text{Inv-Gamma}(\alpha_3, \beta_3), \\ \sigma_d^2 &\sim \text{Inv-Gamma}(\alpha_4, \beta_4). \end{aligned}$$

All parameters in the second-layer priors,

$$(\mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2, \sigma_3^2, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta_1, \beta_2, \beta_3, \beta_4),$$

are assigned in a reasonable way. Furthermore, we also assume that all the priors are independent.

More specifically, the CI method proceeds as follows. There are J items in the test, and each of the J items in the test is examined one at a time. For item j, a size α test of $\delta_j = 0$ is constructed. Let item j follow Eq. (3) and the other items follow Eq. (1). That is, item j is tested if the responses of the focal group need an additional parameter δ_j. The Bayesian analysis via the Markov chain Monte Carlo (MCMC) scheme is implemented to construct the equal-tailed $1 - \alpha$ credible interval for the parameter δ_j. If the interval includes 0, then $\delta_j = 0$ is not rejected. Otherwise, $\delta_j = 0$ is rejected, and hence item j is considered a DIF item.

2.2 The OR Method

The OR method was proposed by Jin, Chen, and Wang (2018) to detect uniform DIF. Let n_{R1j} and n_{R0j} be the numbers of examinees for the reference group who answer item j correctly and incorrectly, respectively; and let n_{F1j} and n_{F0j} be the numbers of examinees for the focal group who answer item j correctly and incorrectly, respectively. For item j, let $\hat{\lambda}_j$ denote the logarithm of the OR of success over failure for the reference and focal groups:

$$\hat{\lambda}_j = \log\left(\frac{n_{R1j}/n_{R0j}}{n_{F1j}/n_{F0j}}\right), \tag{4}$$

which follows a normal distribution asymptotically (Agresti, 2002) with mean λ and standard deviation

$$\sigma(\hat{\lambda}_j) = \left(n_{R1j}^{-1} + n_{R0j}^{-1} + n_{F1j}^{-1} + n_{F0j}^{-1}\right)^{1/2}, \tag{5}$$

where λ is the mean ability difference between the reference and focal groups. For each item j, $\hat{\lambda}_j$, $\sigma(\hat{\lambda}_j)$, and $\hat{\lambda}_j \pm z_{\alpha/2} \times \sigma(\hat{\lambda}_j)$ are computed. Then, find the median for $\hat{\lambda}_1, \hat{\lambda}_2, \ldots,$ and $\hat{\lambda}_J$. An item j is flagged as a DIF item if $\hat{\lambda}_j \pm z_{\alpha/2} \times \sigma(\hat{\lambda}_j)$, the $1-\alpha$ confidence interval of item *j,* does not cover the median of $\hat{\lambda}_1, \hat{\lambda}_2, \ldots,$ and $\hat{\lambda}_J$. Note that this method may not work when the number of examinees are very small because the values of $\hat{\lambda}_j$ cannot be computed when any numbers in Eq. (4) is zero. The scale purification procedures can easily be implemented with the OR method; all that is necessary is the precomputation of the sample median based on presumably DIF-free items. See Jin, Chen, and Wang (2018) for the details.

3 Simulation Study

3.1 Design

In this section, the simulation studies were conducted to compare the performance of the CI and OR methods. In each experiment, we simulated a test consisting of 20 items (i.e., $J = 20$). The number of examinee (P) is 1000. Specifically, we were interested in the comparisons based on the five factors, which were also considered in Simulation Study I of Jin et al. (2018). They were (a) equal and unequal sample sizes of the reference and focal groups (500/500 and 800/200), (b) percentages of DIF items (0, 10, 20, 30 and 40%), (c) DIF patterns: balanced and unbalanced, (d) impact (0 and 1), and (e) purification procedure (with or without). Under the balanced DIF conditions, some DIF items favored the reference group and the other items favored the focal group. By contrast, under the unbalanced DIF conditions, all DIF items favored the reference group.

Item responses were generated according to Eq. (3). The true values of difficulty parameters b_j were generated identically and independently from a uniform distribution between -1.5 and 1.5. The true values of examinee ability parameters θ_p for the reference group ($g_p = 0$) were generated from the standard normal distribution. When impact $= 0$, the true values of θ_p for the focal group ($g_p = 1$) were also generated from the standard normal distribution; when impact $= 1$, they were generated from the normal distribution with mean -1 and variance 1. Under the unbalanced DIF conditions, $d_j - b_j = 0.5$ for all DIF items; under the balanced DIF conditions, $d_j - b_j = 0.5$ for the first half of the DIF items and $d_j - b_j = -0.5$ for the second half of the DIF items. We fixed α, the Type-I error of each test, to 0.05.

To construct the credible intervals, we produced 11,000 MCMC draws with the first 1000 draws as burn-in. A total of 100 replications were carried out under each condition. The performance of these two methods was compared in terms of the FPR and TPR. The FPR was the rate that DIF-free items were misclassified as having DIF whereas the TPR was rate that DIF items were correctly classified as having DIF. The averaged FPR across the DIF-free items and averaged TPR across the DIF items for these two methods were reported. Both the OR and CI methods were implemented by using FORTRAN code with IMSL subroutines, and are available upon request.

3.2 Results

The averaged FPR and TPR of two DIF detection methods for equal (500/500) and unequal (800/200) sample sizes list in Tables 1 and 2, respectively. As expected, both methods yielded well-controlled FPR under the no-DIF (0% DIF items) and balanced DIF conditions, although the OR method was slightly conservative. Similar to Jin, Chen, and Wang (2018)'s study, the FPR larger than or equal to 7.5% was defined as the inflated FPR in the present study. Under the unbalanced DIF conditions, the OR method yielded slightly inflated FPR only when tests had 40% or more DIF items. However, the CI method yielded inflated FPR when tests had 20% or more DIF items under the unbalanced DIF conditions. The TPR of the CI method was higher than that of the OR methods under two following conditions: (i) the balanced DIF conditions and (ii) the unbalanced DIF conditions with 10% DIF items. Furthermore, under these two conditions, the ratio of the TPR of the CI method to that of the OR method with scale purification procedure ranged from 1.01 to 1.27, and it was larger for unequal (800/200) sample sizes than that for equal (500/500) sample sizes. When the total sample size is 1000, the TPR for equal (500/500) sample sizes was higher than that for unequal (800/200) sample sizes. In general, both the FPR and TPR increased with the percentages of DIF items. The TPR for the balanced DIF was higher than that for the unbalanced DIF, except for the OR method when Impact = 0 with equal (500/500) sample size. In general, the TPR was higher when Impact = 0 than that when Impact = 1. The purification procedure increased the TPR for the unbalanced DIF condition, and the higher the percentage of the DIF items, the higher the ratio of the TPR of the OR method with scale purification to that of the OR method without scale purification. By contrast, the purification procedure did not increase the TPR for the balanced DIF condition.

4 Application

In this section, the CI and OR methods described in the previous sections were 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). Each examinee was required to answer 26 questions within 80 min. The 26 questions were further divided into three parts. The total score was 100, and the test was administered under the formula-scoring directions. For the first part, there were 20 multiple-choice questions, and the examinees had to choose one correct answer out of 5 possible choices. For each correct answer, 3 points were granted, and 3/4 point was deducted from the raw score for each incorrect answer. The second part consisted of 4 multiple-response questions, and each question consisted of 5 choices, examinees needed to select all the answer choices that apply. The choices in each item were knowledge-related, but were answered and graded separately. For each correct choice, 1 point was earned, and for each incorrect choice 1 point was deducted from the raw score. The final adjusted scores for each of these two parts started from 0. The last part consisted of 2 calculation problems, and deserved 20 points in total.

The data from 1000 randomly sampled examinees contained the original responses and nonresponses information, but we treated both nonresponses and incorrect answers the same way and coded them as $Y_{pj} = 0$ as Chang et al. (2014) suggested. As for the calculation part, the response Y_{pj} was coded as 1 whenever the original score was more than 7.5 out of 10 points, and zero otherwise (see also Chang et al., 2014). Here, we considered male and female as the reference and focal groups, respectively. Among the 1000 examinees, 692 of them were male and the others were female.

We made more MCMC draws than that in Sect. 3. Specifically, we produced 40,000 MCMC draws with the first 10,000 draws as burn-in. Then we tested $\delta_j = 0$, for $j = 1, \ldots, 26$. Again, we considered $\alpha = 0.05$. The intervals of $\hat{\lambda}_j \pm z_{\alpha/2} \times \sigma(\hat{\lambda}_j)$ for the OR method, which were the same for both with and without purification, and the credible intervals obtained from the real data were summarized in Table 3. Note that the median of $\hat{\lambda}_1, \hat{\lambda}_2, \ldots,$ and $\hat{\lambda}_J$ before and after purification were 0.5687 and 0.6163, respectively, so the OR method identified Items 3, 5, 8, 19 and 23 as DIF items, which were underlined and bolded in Table 3. Table 3 also showed that the

Table 1 Averaged FPR (%) and TPR (%) under the conditions with sample sizes of the reference and the focal groups: 500/500

			FPR			TPR		
Impact	DIF type	DIF (%)	OR with purification	OR without purification	CI	OR with purification	OR without purification	CI
Impact = 0		0	3.35	3.15	5.05	–	–	–
	Balanced DIF	10	3.66	3.22	5.06	85.00	85.50	93.50
		20	3.94	3.63	5.44	84.75	85.25	93.75
		30	3.71	3.43	5.43	86.67	86.67	92.17
		40	4.33	3.33	5.58	88.00	89.25	92.00
	Unbalanced DIF	10	4.06	3.67	5.83	87.00	87.00	88.00
		20	4.63	4.44	**10.56**	87.25	83.00	80.25
		30	4.71	5.50	**17.43**	84.33	76.33	70.67
		40	**7.83**	**10.58**	**29.00**	76.50	63.25	56.50
Impact=1		0	3.10	2.95	5.35	–	–	–
	Balanced DIF	10	2.89	2.83	5.39	81.00	81.00	89.50
		20	3.13	2.94	5.38	86.25	86.75	89.25
		30	3.36	3.00	5.29	84.17	85.50	89.33
		40	3.50	2.92	5.33	82.88	85.50	90.50
	Unbalanced DIF	10	3.28	3.11	6.28	78.00	77.00	80.50
		20	4.50	4.13	**10.38**	77.00	73.00	73.50
		30	5.07	6.64	**15.86**	72.67	65.50	65.33
		40	**9.17**	**11.75**	**27.08**	66.75	55.25	53.13

Note Inflated FPR ($\geq$7.5%) are underlined and bolded

Table 2 Averaged FPR (%) and TPR (%) under the conditions with sample sizes of the reference and the focal groups: 800/200

			FPR			TPR		
Impact	DIF type	DIF (%)	OR with purification	OR without purification	CI	OR with purification	OR without purification	CI
Impact = 0		0	4.35	4.05	4.95	–	–	–
	Balanced DIF	10	4.22	4.00	5.11	67.50	68.50	76.00
		20	4.25	4.00	5.06	68.75	69.75	78.25
		30	4.50	4.21	5.43	67.67	68.83	79.50
		40	4.00	3.67	5.33	68.00	68.88	80.00
	Unbalanced DIF	10	4.50	4.33	5.56	65.50	64.50	70.50
		20	4.25	4.56	**8.38**	60.25	55.75	59.75
		30	6.36	6.29	**12.93**	53.50	47.00	49.00
		40	**10.67**	**11.00**	**19.08**	44.88	36.38	38.25
Impact=1		0	3.55	3.45	5.35	–	–	–
	Balanced DIF	10	3.61	3.38	5.11	60.00	60.00	76.00
		20	4.19	3.50	5.06	63.75	63.75	76.25
		30	4.07	3.50	5.36	63.83	64.50	74.67
		40	5.33	3.50	5.67	65.88	67.13	72.38
	Unbalanced DIF	10	4.00	3.78	6.11	53.00	51.50	66.50
		20	5.13	4.75	**7.75**	54.50	50.50	54.50
		30	6.79	6.50	**11.93**	46.83	40.83	44.50
		40	**14.42**	**12.00**	**16.50**	36.75	29.75	32.88

Note Inflated FPR ($\geq$7.5%) are underlined and bolded

Table 3 The intervals of the OR and CI methods for the real data

Item no.	OR	CI
1	(0.195, 0.7438)	(−0.247, 0.372)
2	(0.391, 1.102)	(−0.650, 0.121)
3	**(−0.396, 0.399)**	**(0.196, 1.034)**
4	(0.233, 1.012)	(−0.481, 0.357)
5	**(−0.203, 0.426)**	**(0.138, 0.846)**
6	(0.567, 1.112)	**(−0.738, −0.111)**
7	(0.377, 0.993)	(−0.564, 0.122)
8	**(−0.168, 0.454)**	**(0.111, 0.812)**
9	(0.312, 0.860)	(−0.404, 0.214)
10	(0.484, 1.210)	*(−0.783, −0.001)*
11	(0.296, 0.850)	(−0.396, 0.232)
12	(0.403, 0.993)	(−0.570, 0.100)
13	(0.219, 0.910)	(−0.431, 0.335)
14	(0.374, 0.925)	(−0.494, 0.127)
15	(0.135, 0.736)	(−0.243, 0.426)
16	(0.168, 0.717)	(−0.228, 0.394)
17	(0.459, 1.246)	(−0.798, 0.044)
18	(0.523, 1.235)	*(−0.829, −0.027)*
19	**(−0.261, 0.300)**	**(0.305, 0.942)**
20	(0.125, 0.677)	(−0.180, 0.447)
21	(0.345, 0.888)	(−0.445, 0.166)
22	(0.193, 0.858)	(−0.365, 0.362)
23	**(−0.164, 0.421)**	**(0.158, 0.804)**
24	(−0.339, 1.256)	(−0.770, 0.855)
25	(0.529, 2.395)	**(−1.915, −0.052)**
26	(−0.149, 1.210)	(−0.700, 0.734)

CI method identified not only Items 3, 5, 8, 19 and 23 as DIF items, but also Items 6, 10, 18 and 25. Based on the result from the OR method, the real data could be contaminated with unbalanced DIF items because the intervals of the identified DIF items all fell on the same side of the median. According to the simulation results in Tables 1 and 2, the CI method yielded inflated FPR when test had 20% or more unbalanced DIF items.

To reduce the inflated FPR of the CI method, we proposed a two-stage CI method to detect DIF items, which was implemented as follows. At the first stage, we detected the DIF items by using the CI method. Suppose $\{i_1, i_2, \ldots, i_k\}$ were the collection of the DIF items identified by the CI method. At the second stage, we check, for $j = 1$, …, k, if item i_k is a real DIF item by deleting the other DIF items, and use only item i_k and the other non-DIF items to fit the Rasch model and then to detect if item i_k is a DIF item based on the CI method again. Based on the two-stage CI method, the identified DIF items were Items 3, 5, 6, 8, 19, 23 and 25, the credible intervals of these items were underlined and bolded in Table 3. Items 10 and 18 were identified as DIF items at the first stage, but were not identified as DIF items at the second stage, and the credible intervals of these two items were marked in *italic* and underlined in Table 3.

5 Concluding Remarks

In this article, we compared the finite sample performance of the CI and OR methods for detecting the need of an additional difficulty parameter for the responses of the focal group when the data follow the Rasch model. Simulation studies showed that the CI method worked better than the OR method under the balanced DIF conditions. However, the CI method yielded inflated FPR under the unbalanced DIF condition. The two methods were then applied to an empirical example. Comparisons of these two methods to other IRT models will be an interesting future line of research.

Acknowledgements The research was supported by Academia Sinica and the Ministry of Science and Technology of the Republic of China under grant number MOST 106-2118-M-001-003-MY2. The authors would like to thank Ms. Yi-Jhen Wu for her helpful comments and suggestions.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York, NY: Wiley.

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. https://doi.org/10.1007/978-3-319-38759-8_3.

Chang, Y.-W., Tsai, R.-C., & Hsu, N.-J. (2014). A speeded item response model: Leave the harder till later. *Psychometrika, 79,* 255–274. https://doi.org/10.1007/s11336-013-9336-2.

Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6,* 269–279. https://doi.org/10.1207/s15324818ame0604_2.

Cohen, A. S., Kim, S. H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20,* 15–26. https://doi.org/10.1177/014662169602000102.

Frederickx, S., Tuerlinckx, F., de Boeck, P., & Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement, 47,* 432–457. https://doi.org/10.1111/j.1745-3984.2010.00122.x.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement, 67,* 373–393. https://doi.org/10.1177/0013164406294781.

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, 8,* 647–667.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum.

Jin, K. -Y., Chen, H. -F., & Wang, W. -C. (2018). Using odds ratios to detect differential item functioning. *Applied Psychological Measurement, 42,* 613–629. https://doi.org/10.1177/0146621618762738.

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement, 75,* 22–56. https://doi.org/10.1177/0013164414529792.

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

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42,* 847–862. https://doi.org/10.3758/brm.42.3.847.

Narayanon, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20,* 257–274. https://doi.org/10.1177/014662169602000306.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 495–502. https://doi.org/10.1007/bf02294403.

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

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17,* 105–116. https://doi.org/10.1177/014662169301700201.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58,* 159–194. https://doi.org/10.1007/bf02294572.

Su, Y.-H., Chang, J., & Tsai, H. (2018). Using credible intervals to detect differential item functioning in IRT Models. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology research* (Vol. 233, pp. 297–304). Switzerland: Springer. https://doi.org/10.1007/978-3-319-77249-3_25.

Psychometric Properties of the Highest and the Super Composite Scores

Dongmei Li

Abstract For students who took college admissions tests multiple times, institutions may have different policies of utilizing the multiple sets of test scores for decision making. For example, some may use the most recent, and others may use the average, the highest, or even the super composite scores by combining the highest subject test scores from each administration. Previous research on these different score use policies mainly focused on their predictive validity with little discussion about their psychometric properties. Through both theoretical and empirical investigations, this study showed how the bias, the standard error of measurement, and the reliability of scores for these different policies compare with each other and how these properties change for each score type as the number of test events increased.

Keywords Super composite · Reliability · Standard error of measurement · Sample maxima

1 Introduction

Academic achievement test programs often report scores on each subject test as well as a composite score based on all the subject test scores. For tests that are used for college admission purposes, students may choose to take the test multiple times if retesting is allowed. Therefore, institutions may receive score reports for a student from multiple test administrations. Different institutions may have different policies regarding how to use these multiple scores for decision making. Whereas some might choose to use the most recent or the average of the scores from different administrations, others might choose to use the highest reported composite scores or even the super composite scores, which are obtained by re-calculating the composite scores using the highest subject test scores from each test event.

There is no consensus in the literature (Boldt, 1977; Boldt, Gentra, & Courtney, 1986; Linn, 1977; Mattern, Radunzel, Bertling, & Ho, 2018; Patterson, Mattern, &

D. Li (✉)
ACT, Inc., 500 ACT Drive, Iowa City, IA 52243, USA
e-mail: dongmei.li@act.org

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_29

Swerdzewski, 2012; Roszkowski & Spreat, 2016) regarding which is the best way to treat scores from multiple test administrations for college admission purposes. Whereas the average score was often found to have slightly higher correlation with college grades than the other scoring methods, it was also found to have the greatest extent of underprediction of the outcome variable (Boldt, Centra, & Courtney, 1986). Though there are cautions against their potential to maximize positive measurement error (e.g. ACT, 2018), super composite scores were also found to slightly better predict college first-year grade point average (FYGPA) than other scoring methods (Mattern, Radunzel, Bertling, & Ho, 2018). The use of super composite scores are also more controversial than others because it causes more fairness concerns if students do not have the same opportunity to retest.

Most of the previous research about the different scoring policies focused on their predictive validity. Little can be found in the literature about the psychometric properties of the scores resulting from these different scoring polices, especially for the highest and the super composite scores. Though it is a standard practice for test programs to report the standard error of measurement (SEM) and the reliability of reported test scores, these psychometric properties will change when decision making is based on scores from multiple test administrations, whether by taking the average or the highest or a combination of the highest of each subject test. Furthermore, psychometric properties of test scores are known to affect their predictive validity. For example, the predictive strength of a variable will decrease as its reliability gets lower due to the weakening effect of measurement error (Spearman, 1904). Therefore, a better understanding of the psychometric properties of scores from these different scoring policies is needed in order to inform the evaluation of these approaches. The purpose of this study is to investigate the psychometric properties of the highest scores and the super composite scores and compare them with those of the average scores and scores from a single administration.

Four approaches of utilizing multiple sets of scores were investigated in the study: the most recent, the average, the highest, and the super composite scores. Their psychometric properties, including systematic error (i.e. bias), SEM, and reliability, were investigated both theoretically and empirically. The theoretical investigation was done under the assumptions of classical test theory utilizing statistical properties of sample means and sample maxima. The purpose of the theoretical investigation was to derive relationships that can be used to predict psychometric properties of the highest and the super composite scores based on the reported SEM or reliability, which are meant to apply only to scores based on a single test administration. A simulation study was then conducted to empirically compare these properties for the four types of scores and to confirm the relationships derived from the theoretical investigation. Both the theoretical and the empirical investigations in this study were conducted under the assumptions of classical test theory and also under the assumption that no real ability changes occurred across the different test events.

2 Theoretical Investigation

To facilitate discussions, the four types of composites are defined in more detail through mathematical expressions, and measurement error and reliability are then described under the assumptions of classical test theory. After that, the expected values and variances of sample means and sample maxima are used to derive relationships that can be used to predict the psychometric properties of scores from the different scoring policies.

2.1 Mathematical Expressions of the Four Types of Composites

Let $C_1, C_2, \ldots\ldots C_M$ represent composite scores from test administrations 1 through *M*, each composite score from administration $m(C_m)$ being a function (*f*) of the *N* subject test scores represented by $S_{1m}, S_{2m}, \ldots\ldots S_{Nm}$, that is,

$$C_m = f(S_{1m}, S_{2m}, \ldots\ldots S_{Nm}). \tag{1}$$

Let C_{rec}, C_{avg}, C_{max}, C_{sup} represent the most recent, the average, the highest, and the super composite score from *M* test administrations. These scores can be expressed as

$$C_{rec} = C_M = f(S_{1M}, S_{2M}, \ldots\ldots S_{NM}), \tag{2}$$

$$C_{avg} = \frac{1}{M}\left(\sum\nolimits_{m=1}^{M} C_m\right) = \frac{1}{M}\left(\sum\nolimits_{m=1}^{M} f(S_{1m}, S_{2m}, \ldots\ldots S_{Nm}\right), \tag{3}$$

$$C_{max} = max(C_1, C_2, \ldots\ldots C_M), \text{ and} \tag{4}$$

$$\begin{aligned} C_{sup} = f(&max(S_{11}, S_{12}, \ldots\ldots S_{1M}), max(S_{21}, S_{22}, \ldots\ldots S_{2M}), \ldots\ldots, \\ &max(S_{N1}, S_{N2}, \ldots\ldots S_{NM})). \end{aligned} \tag{5}$$

To make it simple, the composite score in this study is defined as the average of all the subject test scores. That is, the composite score from a test administration *m* is

$$C_m = f(S_{1m}, S_{2m}, \ldots\ldots S_{Nm}) = \frac{\sum_1^N S_{nm}}{N}. \tag{6}$$

2.2 Measurement Error and Reliability Under Classical Test Theory Assumptions

Classical test theory (Lord & Novick, 1968) assumes that all observed test scores (X) are composed of true scores (T) and errors (E) (i.e. X = T + E), where true scores are defined as the expected value of observed scores over an infinite number of replications. Errors are assumed to be independent of true scores, and the expected value of errors across an infinite number of replications is 0. Therefore, the observed score variance from a population of examinees is composed of true score variance and error variance, which can be expressed as

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2. \tag{7}$$

Classical test theory defines reliability as the correlation of observed scores between two parallel test forms (*denoted as* $\rho_{XX'}$), which can be calculated either as the squared correlation between true and observed scores or as the ratio of true score variance and observed score variance, i.e.,

$$\rho_{XX'} = \rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}. \tag{8}$$

2.3 Distributions of Sample Means and Psychometric Properties of the Average Scores

Psychometric properties of the different statistics of multiple scores can be derived using properties of the sampling distributions of these statistics. In fact, the statistical properties of sample means are well known and have long been used in the derivation of psychometric properties of mean scores, though they are seldom used to discuss such properties of the average composite scores in the context of admission decisions with multiple sets of scores. This section shows how the well-known statistical properties of sample means can be used to understand the psychometric properties of the average composite scores, which is not new except for the application in this context but will be helpful for discussions in the next section about the highest and the super composite scores.

For a random sample of size *n* from the distribution of a variable X, the expected value of the sample mean $\overline{X}$ is

$$\boldsymbol{E}\left(\overline{X}\right) = \boldsymbol{E}(X), \tag{9}$$

and its variance is

$$\sigma^2_{\overline{X}} = \frac{\sigma^2_X}{n}. \tag{10}$$

Applying Eq. (9) in the context of the expected value of measurement error for the average composite score as defined in Eq. (3), it follows that the average is an unbiased estimate of the true scores because its expected value of measurement error is

$$\boldsymbol{E}\left(E_{avg}\right) = \boldsymbol{E}\left(\overline{E}\right) = \boldsymbol{E}(E) = 0. \tag{11}$$

Applying Eq. (10) in the context of error variance, the error variance of the average composite score is

$$\sigma^2_{E_{avg}} = \frac{\sigma^2_E}{n}. \tag{12}$$

Therefore, the reliability of the average composite score can be expressed as

$$\rho_{XX'_{avg}} = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_{E_{avg}}} = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_E/n}. \tag{13}$$

Dividing the denominator and numerator of the last part of Eq. (13) by σ^2_X, it can be easily proven that the reliability of average scores can be expressed as

$$\rho_{XX'_{avg}} = \frac{n\rho_{XX'}}{1 + (n-1)\rho_{XX'}}, \tag{14}$$

which is the Spearman-Brown formula and allows the prediction of reliability for the average composite scores across n test events based on the known reliability $\rho_{XX'}$ reported for the composite scores of the test.

The next section shows that utilizing the statistical properties of sample maxima, similar predictions can be made for the highest and the super composite scores.

2.4 Distributions of Sample Maxima and Psychometric Properties of the Highest and the Super Composite Scores

The statistical properties of sample maxima are far less well known for psychometric researchers than those of sample means because they often involve integration and cannot be expressed in a nice simple formula. Probably that is one of the reasons that the psychometric properties of the highest or the super composite scores are seldom discussed. Fortunately, Chen and Tyler (1999) provided an accurate approximation for the expected values and the standard deviations of the maxima of samples from standard normal distributions. They showed that the expected value of the sample maxima can be accurately approximated by the expression $\Phi^{-1}(0.5264^{1/n})$, where n is the sample size and Φ^{-1} is the inverse of the Gaussian cumulative distribution function, and that the standard deviation of sample maxima can be approximated by the expression $0.5[\Phi^{-1}(0.8832^{1/n}) - \Phi^{-1}(0.2142^{1/n})]$.

Applying these approximations in the context of measurement error, which is assumed to be normally distributed with a mean of 0 and a standard deviation of σ_E for the composite scores, or equivalently if it is assumed that the measurement errors for the subject test scores are uncorrelated and normally distributed with a mean of 0 and a standard deviation of $\sigma_{E_{s_i}}$, then the expected value of measurement error for the highest scores can be expressed by

$$E(E_{max}) = \sigma_E f(n) = f(n)\frac{\sqrt{\sum_i^N \sigma_{E_{s_i}}^2}}{N}, \tag{15}$$

where

$$f(n) = \Phi^{-1}\left(0.5264^{\frac{1}{n}}\right), \tag{16}$$

and the error variance of the highest scores is

$$\sigma_{E_{max}}^2 = \sigma_E^2[g(n)]^2, \tag{17}$$

where

$$g(n) = 0.5\left[\Phi^{-1}\left(0.8832^{\frac{1}{n}}\right) - \Phi^{-1}\left(0.2142^{\frac{1}{n}}\right)\right]. \tag{18}$$

Therefore, the reliability of the highest composite score is

$$\rho_{XX'_{max}} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_{E_{max}}^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2[g(n)]^2}. \tag{19}$$

By diving the denominator and numerator of the last part of Eq. (19) by σ_X^2, it can be easily proven that the reliability of the highest scores can be predicted by

$$\rho_{XX'_{max}} = \frac{\rho_{XX'}}{\rho_{XX'} + (1 - \rho_{XX'})[g(n)]^2}. \tag{20}$$

Equation (20) in combination with Eq. (18) can be used to predict the reliability of the highest composite scores based on any number of administrations given the reported reliability of composite scores for a test.

For the super composite scores as defined in Eq. (5), the expected value of measurement errors (i.e. bias) is

$$\boldsymbol{E}\left(E_{sup}\right) = f(n)\frac{\sum_{i}^{N}\sigma_{E_{s_i}}}{N}. \tag{21}$$

When composite scores are defined as the average of the component tests, the error variance of super composite scores is

$$\sigma_{E_{sup}}^2 = \frac{\sum_1^N \sigma_{E_{max S_i}}^2}{N^2} = \frac{\sum_1^N \sigma_{E_{S_i}}^2 [g(n)]^2}{N^2} = \sigma_E^2 [g(n)]^2 = \sigma_{E_{max}}^2. \tag{22}$$

The reliability of the super composite scores is then

$$\rho_{XX'_{sup}} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_{E_{sup}}^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_{E_{max}}^2} = \rho_{XX'_{max}}. \tag{23}$$

By comparing Eq. (21) with Eq. (15), it can be shown that the expected error of super composite scores is higher than that of the highest composite scores. That is to say, both the highest and the super composite scores are biased estimates of true composite scores, but the super composite scores are more biased than the highest composite scores. Equations (22) and (23) show that the error variance of super composite scores and that of the highest composite scores are mathematically equal to each other, under the condition that composite scores are defined as the average of the subject test scores. That is to say, when composite scores are defined as the simple average of the subject test scores, Eq. (20) in combination with Eq. (18) can also be used to predict the reliability of super composite scores based on any number of administrations. These equations also show that as the number of test events increases, the bias and reliability will both increase for the highest and the super composite scores. Whereas bias increases faster for super composite scores than for the highest scores, reliability increases at the same rate for the highest and super composite scores.

3 Empirical Investigation

3.1 Data Simulation

To illustrate the relationships discussed above, true scores for four subject tests were generated for 10,000 examinees from a multivariate normal distribution based on means, variances, and covariances of the ACT test (ACT, 2014). Measurement errors for 10 test events were generated from normal distributions with a mean of 0 and a standard deviation of 2. The composite score was defined as the simple average of the four subject test scores. The true composite score is the average of the four true subject test scores. Observed scores for each examinee from each test event were calculated as the sum of the true score and the measurement error. Then scores based on the four different scoring policies (the most recent, the average, the highest, and the super composite scores) were calculated for each examinee based on 1 through 10 test events, respectively.

Two sets of observed scores were used for all analysis: one was the unrounded scores generated above and the other is the rounded and truncated versions of the above scores to mimic normal score reporting. Following the practice of the ACT test, observed scores were rounded to whole numbers and truncated to be within ranges of 1 to 36 for the four subject tests and for the composite scores. The reason for doing this is that some of the predictions may work well only for the unrounded and untruncated scores.

3.2 Indices of Psychometric Properties

The following statistics were calculated for scores based on each scoring policy for both the unrounded and the rounded and truncated scores: (1) bias, the overall mean difference between the observed composite scores and the true composite scores across all examinees; (2) SEM, the standard deviation of the differences between each observed composite scores and the true composite scores across all examinees; and (3) reliability defined as the squared correlation of the observed composite scores and the true composite scores as shown in Eq. (8).

Meanwhile, these statistics were predicted for the average, the highest, and the super composite scores based on relevant equations presented in Sects. 2.3 and 2.4. Equation (15) in combination with Eq. (16) were used to calculate the predicted bias for the highest composite scores, and Eq. (21) in combination with Eq. (16) were used to predict the expected bias for the super composite scores. Equation (17) in combinations with Eq. (18) were used to calculate the predicted values of the SEMs of both the highest and the super composite scores, and Eq. (12) was used to predict the SEMs of the average composite scores. Equation (20) in combination with Eq. (18) were used to predict the reliability of both the highest and the super composite scores, and Eq. (14) was used to predict the reliability of the average composite scores.

3.3 Results

Results on bias, SEM, and reliability are presented in Figs. 1, 2, and 3, respectively. In these figures, the horizontal axes represent the number of test events, and the horizontal axes represent the statistic of interest. Each figure is composed of two parts: Figure (a) presents results calculated from the simulated data, and Figure (b) overlays curves representing the predicted values upon the curves shown in Figure (a). When the predictions are accurate, the curves representing the calculated values and those representing the predicted values overlap.

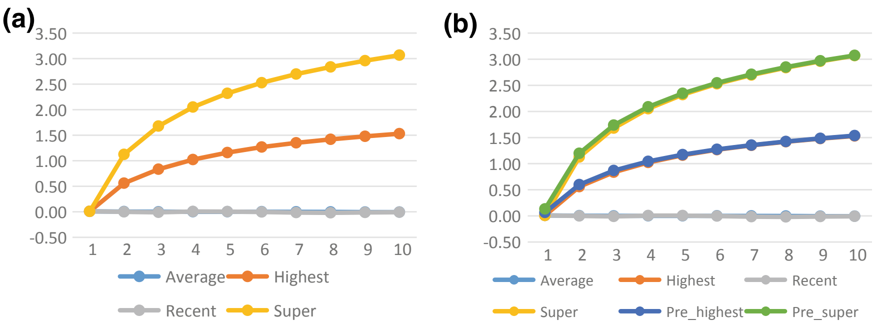


Fig. 1 Bias of various scoring type for different numbers of test events

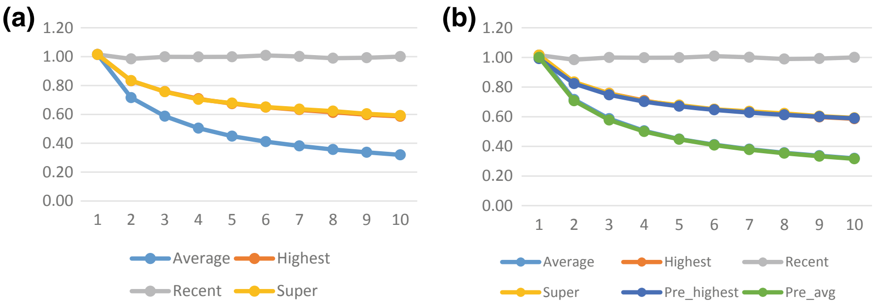


Fig. 2 SEM of various scoring type for different numbers of test events

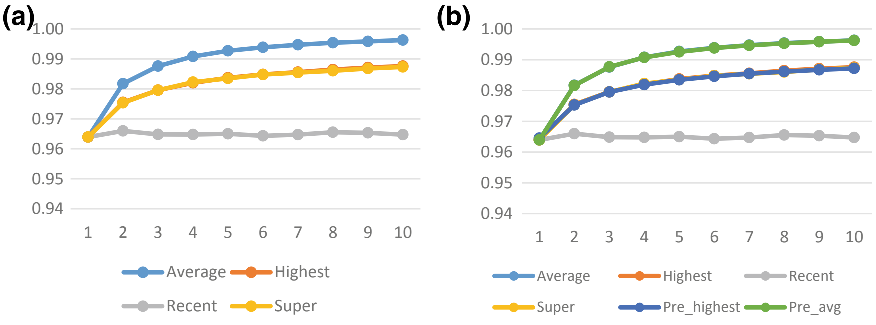


Fig. 3 Reliability of various scoring types for different numbers of test events

Figure 1 shows the bias of the four types of composite scores as the number of test events increases. Figure 1a shows their biases calculated based on the simulated data. As expected, biases for the average scores and the most recent scores were both 0 regardless of the number of test events, so the lines of these two scores overlap. Both the highest and the super composite scores had positive biases thus overestimated the true composite. The extent of their overestimation increased as the number of test events increased, but the super composite scores overestimated to a greater extent than the highest composite. Figure 1b added to what has been shown in Fig. 1a the predicted bias for the highest scores (Pre_highest) and that for the super composite scores (Pre_super). The fact that the curves for the predicted values overlap with the curves calculated from data indicates that the predictions were very accurate.

Figure 2 presents the SEMs of the four types of composite scores. Figure 2a, based on results calculated from data, shows that SEMs for the average are the smallest of the four, and that the SEMs for the most recent are the largest. The SEMs for the highest and the super composite scores are the same with overlapping curves, and they are higher than those of the average scores but lower than the most recent. Figure 2b added two series to those in Fig. 2a—the predicted SEMs for the highest (Pre_highest), which are equal to those of the super composite, and the predicted SEMs for the average (Pre_avg). The curves for the predicted values overlapped with the curves calculated based on empirical data, indicating that the predictions were accurate.

Figure 3 presents the reliability for the four score types as the number of test events increases. Figure 3a, based on results calculated from the simulated data, shows that the average composite had the highest reliability, followed by the highest and super composite with overlapping curves. The results being the same for the highest and the super composite scores were expected because of how the composite score was defined. The most recent had the lowest reliability among the four. Figure 3b added the predicted reliability for the highest and the super composite scores (Pre_highest) as well as that for the average (Pre_avg). Again, the prediction worked very well.

3.4 Effects of Rounding and Truncation

As mentioned earlier, all the above analyses were also conducted for the integer test scores obtained through rounding and truncating the continuous scores to whole numbers within a fixed range. Results showed that the rounding and truncation did affect the SEM and reliability of all four types of scores, though their bias was not affected. The SEM for the integer scores is higher and the reliability is lower than that of the continuous scores. Though the predictions worked almost perfectly for the unrounded and untruncated scores, they underpredicted SEM and overpredicted reliability of the average or the highest and the super composite scores when rounding and truncation were involved. Figure 4 shows the reliability estimates based on the integer scores calculated from the simulated data, and the reliability predictions for the average and the highest scores. A preliminary follow up research showed that

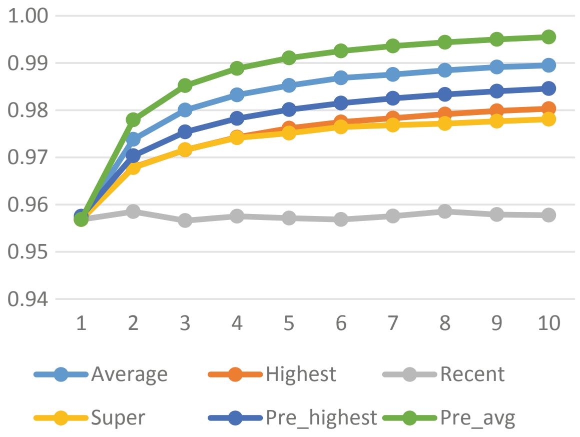


Fig. 4 Reliability of various types of scoring when scores are rounded and truncated

truncation affected the range of observed scores and caused errors to be negatively correlated with true scores, which may be a reason for the diminished reliability of the truncated scores. However, more research is needed to understand exactly how the rounding and truncation process affects measurement error and reliability.

4 Conclusion and Discussion

Through theoretical and empirical investigations, this study has shown that when no real ability changes occur between testing, the highest and the super composite scores are biased estimates of true scores and they tend to overestimate more as the number of test events increases. However, both the highest and the super composite scores are more reliable than the most recent when scores from multiple test events are taken into account. The study also shows that though super composite scores overestimate to a larger extent than the highest composite scores, their SEM and reliability are both equivalent when composite score is a simple average of the component scores. Furthermore, the study not only showed that the Spearman-Brown formula can be used to predict reliability of the average composite scores but also derived formulas to predict the bias, the SEM, and the reliability of the highest and the super composite scores. The predictions worked very well when scores were not rounded or truncated, but rounding and truncation was shown to have an impact on SEM and reliability which needs further study.

Findings from this study have practical implications. First, the impact of bias of the highest and the super composite scores should be taken into consideration to ensure fairness in decision making. If all students were tested the same number of times, then scores for all would have been inflated to the same extent, then fairness is not threatened. However, if students have different opportunities for re-testing, then those who tested more times could be potentially advantaged when decisions are based on the highest and especially the super composite scores. Second, the fact that the highest and the super composite scores do have higher reliability than the most recent indicates that the highest or super composite scores may have stronger predictive power than the most recent, though lower than the average. Therefore, higher correlations with criterion scores could be observed for the highest and super composite scores than the most recent, as has been shown in some earlier referenced studies. However, because the highest and the super composite scores are positively biased estimates of true scores, predicted criterion scores with these scores through linear regression are expected to be lower than that based on the recent scores. When a different pattern is found in empirical studies with real data (e.g. Mattern, Radunzel, Bertling, & Ho, 2018), it is a strong indication that factors other than measurement error are involved.

Results from this study should be interpreted keeping in mind the assumptions and the limitations of the study. One limitation was that the study did not take into account potential real ability changes across test events, and only showed results when everyone had the same number of test events. The reality of testing is much more complicated. For example, it is often not known who would choose to retest and why they retested. Institutions need to make decisions for students potentially with different opportunities, different motivations, and different reasons for retesting. Nevertheless, the psychometric properties revealed in this study can provide useful information in the evaluation of different score use policies.

References

ACT (2014). *The ACT® Technical Manual*. Iowa City, IA: ACT, Author.

ACT (2018). The *ACT® Test User Handbook for Educators*. Iowa City, IA: Author.

Boldt, R. F. (1977). *Evaluation of three methods for treating repeaters' scores*. Princeton, NJ: Law School Admission Council.

Boldt, R. F., Centra, J. A., & Courtney, R. G. (1986). *The validity of various methods of treating multiple SAT® scores*. New York, NY: The College Board.

Chen, C., & Tyler, C. (1999). Accurate approximation to the extreme order statistics of Gaussian samples. *Communications in Statistics—Simulation and Computation, 28*(1), 177–188.

Linn, R. L. (1977). *On the treatment of multiple scores for law school admission test repeaters*. Princeton, NJ: Law School Admission Council.

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

Mattern, K., Radunzel, J., Bertling, M., & Ho, A. (2018). How should colleges treat multiple admissions test scores? *Educational Measurement: Issues and Practice, 37*(3), 11–23.

Patterson, B., Mattern, K., & Swerdzewski, P. (2012). Are the best scores the best scores for predicting college success? *Journal of College Admission, 217,* 34–45.

Roszkowski, M., & Spreat, S. (2016). Retaking the SAT may boost scores but this doesn't hurt validity. *Journal of the National College Testing Association, 2*(1), 1–16.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72–101.

A New Equating Method Through Latent Variables

Inés Varas, Jorge González and Fernando A. Quintana

Abstract Comparability of measurements is an important practice in different fields. In educational measurement, equating methods are used to achieve the goal of having comparable scores from different test forms. Equated scores are obtained using the *equating transformation* which maps the scores on the scale of one test form into their equivalents on the scale of another for the case of sum scores. Such transformation has been typically computed using continuous approximations of the score distributions, leading to equated scores that are not necessarily defined on the original discrete scale. Considering scores as ordinal random variables, we propose a latent variable formulation based on a flexible Bayesian nonparametric model to perform an equipercentile-like equating that is capable to produce equated scores on the original discrete scale. The performance of our model is assessed using simulated data under the equivalent groups equating design. The results show that the proposed method has better performance with respect to a discrete version of estimated equated scores from traditional equating methods.

Keywords Test equating · Latent variable representation · Bayesian nonparametric model

I. Varas (✉) · J. González · F. A. Quintana
Departamento de Estadística, Pontificia Universidad Católica de Chile, Santiago, Chile
e-mail: imvaras@mat.uc.cl

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

F. A. Quintana
e-mail: quintana@mat.uc.cl

I. Varas · J. González
Laboratorio Interdisciplinario de Estadística Social, LIES, Facultad de Matemáticas,
Pontificia Universidad Católica de Chile, Santiago, Chile

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_30

1 Introduction

The purpose of test equating methods is to allow the scores on different test forms to be used interchangeably (González & Wiberg, 2017; Kolen & Brennan, 2014; von Davier, Holland, & Thayer, 2004). Let X and Y be two test forms, and X and Y the score random variables defined on sample spaces $\mathscr{X}$ and $\mathscr{Y}$, respectively, which correspond to the score scales (González & Wiberg, 2017). Let us denote by G_X and G_Y the cumulative distribution functions (cdf) associated to each of these random variables.

An equated score on the scale $\mathscr{Y}$ is the result of applying an equating transformation $\varphi : \mathscr{X} \mapsto \mathscr{Y}$ to a score x in $\mathscr{X}$, so that x is mapped into its equivalent, $y^{\star}$, on the scale $\mathscr{Y}$. The commonly used equipercentile equating transformation (Braun & Holland, 1982) is obtained when scores x and $y^{\star}$ are considered to be equated if $G_X(x) = G_Y(y^{\star})$. Using this relation, the equipercentile function becomes

$$\varphi_Y(x) = G_Y^{-1}\left(G_X(x)\right). \tag{1}$$

It is well known in the equating literature that (1) is an ill-defined transformation because most of the times $\mathscr{X}$ and $\mathscr{Y}$ are subsets of the integers number set.[1] With discrete scale scores, both cdfs G_X and G_Y are discrete and $G_X(x)$ will not coincide with $G_Y(y)$ for any possible score on $\mathscr{Y}$. The common approach to solve this problem is to "continuize" the discrete score distributions G_X and G_Y in order to use (1) in a proper way. Examples of equating transformations computed in this way are the percentile-rank and the kernel equating (von Davier, Holland, & Thayer, 2004) transformations, which use linear interpolation and Gaussian kernel smoothing for continuization, respectively.

A common feature of all equating methods based on the continuization of G_X and G_Y is that $y^{\star} \notin \mathscr{Y}$, i.e., equated scores are not integer values anymore. Our approach aims at developing an equating method that respects the discrete nature of test scores and thus gives as a result an equated score value that is properly defined on $\mathscr{Y}$. We propose a model where test scores are modeled as a latent representation of discrete ordinal random variables. The latent representation is based on a flexible Bayesian nonparametric model (Kottas, Muller, & Quintana, 2005). Because the latent variables are continuous, an equipercentile-like equating method as defined in (1) can be performed, and the latent representation used to obtain equated scores on the original scale.

The paper is organized as follows. The latent representation for ordinal variables, the nonparametric Bayesian models, and the latent equating method (LE) are described in Sect. 2. In Sect. 3, the performance of the proposed equating method is evaluated in a simulation study. Conclusions and further work are discussed in Sect. 4.

[1] Equating methods for continuous-type scores, such that those obtained using an IRT model, have been developed but will not be considered in this paper.

2 Latent Modeling Approach

In this section we give a brief description of the latent formulation for ordinal random variables as well as the basics of Bayesian nonparametric models that are used in the proposed method. The new equating method is described at the end of this section.

2.1 *Ordinal Random Variables*

Ordinal categorical variables arise frequently in different fields such as health and social sciences research. In the former, it is of interest to link some factors on the severity of a disease, whereas in the latter, one might be interested, for instance, in analyzing the relationship between the mother's education level and the student's achievement. Different statistical models have been developed for the analysis of this type of variables, most of them in the context of regression models.

Let W be an ordinal random variable with support given by $\{w_0, w_1, \ldots, w_{N_W}\}$ such that $w_0 < w_1 < \cdots < w_{N_W}$. Consider a random sample $W_1, \ldots, W_n$ of W. The latent response formulation (Pearson, 1901) assumes that each response W_i is the partial observation of a continuous latent variable $W_i^\star$. For an ordinal variable this relation can be written as

$$W_i = \begin{cases} w_0 & \text{if} \quad \varrho_0 < W_i^\star \leq \varrho_1 \\ w_1 & \text{if} \quad \varrho_1 < W_i^\star \leq \varrho_2 \\ \vdots & \vdots \qquad\qquad \vdots \\ w_{N_W} & \text{if} \; \varrho_{N_W} < W_i^\star < \varrho_{N_W+1} \end{cases}, \tag{2}$$

where $\{\varrho_0, \varrho_1, \ldots, \varrho_{N_W}, \varrho_{N_W+1}\}$ are thresholds parameters such that $-\infty = \varrho_0 < \varrho_1 < \ldots < \varrho_{N_W} < \varrho_{N_W+1} = +\infty$. In a regression context, $W_i^\star = \eta_i + \epsilon_i$, where η_i is a linear predictor and ϵ_i is an error term. If a normal distribution is assumed for ϵ_i, then the ordinal probit model is obtained (Aitchison & Silvey, 1957; McKelvey & Zavoina, 1975). Instead, if the error term is assumed to have a logistic distribution, the ordinal logit model is obtained (McCullagh, 1980).

Using the representation in (2), it follows that the probability distribution of W_i is specified in terms of the probability distribution of $W_i^\star$, showing that:

$$P(W_i = w_{ik}) = P(W_i^\star \leq \varrho_{k+1}) - P(W_i^\star \leq \varrho_k) \qquad k = 0, \ldots, N_W\,. \tag{3}$$

2.2 *Bayesian Nonparametric Models*

Bayesian nonparametric models are based on random probability measures (RPM) which act as prior probability models defined over distribution functions (Ghosh &

Ramamoorthi, 2003). The Dirichlet process (DP) prior (Ferguson, 1973) is arguably the most used RPM in this framework. A random distribution function F that comes from a DP with mass parameter M and baseline probability measure G_0, written as $F \sim DP(M, G_0)$, can be described using its stick breaking representation (Sethuraman, 1994) in the following way. Let θ_j $(j = 1, 2, \ldots)$ be independent and identically distributed random variables from a distribution function G_0, and ν_i $(i = 1, 2, \ldots)$ independent and identically distributed variables from a beta distribution $Beta(1, M)$. Then, the stick breaking representation of F is written as:

$$F(\cdot) = \sum_{j=1}^{\infty} p_j \delta_{\theta_j}(\cdot), \tag{4}$$

where $p_1 = \nu_1$, $p_j = \nu_j \prod_{i<j}(1 - \nu_i)$, and $\delta(\cdot)$ denotes a point mass at θ_j. From this representation, any realization from a DP is discrete with probability 1 (Blackwell & MacQueen, 1973). Ishwaran and James (2001) proposed several alternative RPMs using the stick breaking representation. One of them is defined by truncating the countable sum in (4) at a truncation level of N terms with $\upsilon_N = 1$ and $p_N = 1 - \sum_{i<N} p_i$. Posterior computations under this prior model are implemented by using the blocked Gibbs sampler algorithm (Ishwaran & James, 2001).

Nonparametric Bayesian models are commonly used in density estimation. However, as a consequence of the discrete nature of distribution functions sampled from a DP, models based on DP priors are no longer useful in the context of continuous density estimation. To deal with this problem, the DP mixture model (DPM) has been proposed as a mixture of a smooth continuous density and a DP prior (Ferguson, 1983; Lo, 1984). Let us consider Z as a continuous random variable with density $f(z)$ defined by a DPM. Then,

$$f(z) = \int p(z \mid \theta) G(d\theta) \tag{5}$$

$$G \sim DP(M, G_0)\ , \tag{6}$$

where for every $\theta \in \Theta$, $p(z \mid \theta)$ is a continuous density function, $\Theta \subset \mathbb{R}^p$ and G is a DP defined on Θ.

The latent formulation for ordinal random variables described in Sect. 2.1 has been useful in density estimation (Shah & Madden, 2004; Ghosh, Burns, Prager, Zhang, & Hui, 2018). Kottas et al. (2005) considered a DPM model for the latent variable $W_i^{\star}$ in the context of modeling multivariate ordinal data. This approach is more convenient than other alternatives because it approximates any probability distribution of ordinal variables. As pointed out by Kottas et al. (2005), this approximation is not based on random thresholds so there is no loss of generality in assuming them to be fixed. From a practical perspective, this means that thresholds need not be estimated, but rather they are considered as fixed known values.

Bayesian nonparametric models have also been used in test equating. Karabatsos and Walker (2009) argued that continuous approximations of the discrete score distributions G_X and G_Y for traditional methods such as the percentile-rank equating, the linear equating, the mean equating (Kolen & Brennan, 2014) and the kernel equating (von Davier, Holland, & Thayer, 2004) methods, are all based on parametric assumptions to build a continuous version of the cdfs of X and Y. These authors proposed a nonparametric Bayesian model for the score distributions using Bernstein polynomials process priors. González, Barrientos, and Quintana (2015) extended this model and considered a nonparametric Bayesian model that allows the use of covariates. However, as it is the case for traditional methods, neither of these two approaches produce equated scores that are properly defined on the original discrete scale. We propose an alternative equating method that tackles this issue by using a latent representation of scores. This method is developed by considering a Bayesian nonparametric model for the latent variable associated to the ordinal score random variable. The use of the latent formulation strategy in density estimation as applied for the estimation of score distributions is explained in the following section.

2.3 Bayesian Nonparametric Latent Approach for Score Distributions

Let $X_1, \ldots, X_{n_X}$ and $Y_1, \ldots, Y_{n_Y}$ be two random samples of sizes n_X and n_Y, respectively. We assume that the scales scores of X and Y are defined by $\mathscr{X} = \{x_0, \ldots, x_{N_X}\}$ and $\mathscr{Y} = \{y_0, \ldots, y_{N_Y}\}$. Because both $\mathscr{X}$ and $\mathscr{Y}$ define an order relation between scores we consider X and Y as ordinal random variables. Under this assumption we develop an equating method based on the latent representation of ordinal variables described in Sect. 2.1.

Before defining the equating method, we describe the proposed model for the scores distributions. We define the model only for X scores but a similar formulation can be made for Y. Let us consider $X_1^\star, \ldots, X_{n_X}^\star$ the latent variable associated to each X_i from the latent formulation (2). Following ideas found in Kottas et al. (2005), we propose a DPM model for the latent variable $X_i^\star$ based on a finite DP prior which considers the number of possible scores of the test X as the truncation level. The proposed model can be written as follows:

$$X_i = x_h \Leftrightarrow \varrho_h < X_i^\star \leq \varrho_{h+1} \; i = 1, \ldots, n_X, \; x_h \in \mathscr{X} \tag{7}$$

$$X_i^\star \mid \theta, K_i \overset{ind}{\sim} N(x_i^\star \mid \mu_{K_i}, 1/\sigma_{K_i}^2) \tag{8}$$

$$K_i \mid \mathbf{p} \overset{iid}{\sim} \sum_{k=1}^{N_X+1} p_k \delta_k(K_i) \tag{9}$$

$$\theta_j \mid \psi \sim N(\mu_j \mid \lambda, \tau/\sigma_j^2) Gamma(\sigma_j^2 \mid \alpha_0, \beta) \tag{10}$$

where $(\varrho_0, \varrho_1, \ldots, \varrho_{N_X})$ is the vector of fixed thresholds, $\theta_j = (\mu_j, \sigma_j^2)$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{K_X+1})$, $\psi = (\lambda, \tau, \beta)$ and $N_X + 1$ is the truncation point of the DP prior. The full model is completed assigning priors on ψ and M, the latter being the parameter that controls the DP prior. In this case we consider:

$$\begin{aligned}\lambda &\sim N(q_0, Q_0)\\ \tau &\sim InvGam(w_0, W_0)\\ \beta &\sim Gamma(c_0, C_0)\\ M &\sim Gamma(a_0, b_0)\ .\end{aligned}$$

Hyperparameters were chosen as in Kottas et al. (2005) and set to $q_0 = 0$, $Q_0 = W_0 = 49$, $\alpha_0 = w_0 = 3$, $c_0 = 4$, $C_0 = 2$ and $a_0 = b_0 = 1$. This choice is motivated due to the fact that marginal prior moments for the parameters are finite and spread distributions for the parameters in the mixing distribution are obtained.

All posterior conditional distributions are readily sampled by the implementation of a blocked Gibbs sampler algorithm (Ishwaran & James, 2001). After L iterations of the algorithm, we obtain posterior samples of all parameters in the model. In particular, in each iteration we obtain samples from the posterior predictive distribution of $X^\star$ which has the following structure

$$F_{X^\star}^{(l)}(x^\star) = \int_{-\infty}^{x^\star} \sum_{k=1}^{N_X+1} p_k^{(l)} N(t \mid \mu_k^{(l)}, 1/\sigma_k^{2(l)})dt\ , \tag{11}$$

where $\{(\mu_k^{(l)}, \sigma_k^{2(l)}), l = 1, \ldots, L\}$ are the sampled parameters from the posterior distribution. As mentioned before, a similar formulation is used to obtain $F_{Y^\star}^{(l)}(y^\star)$.

2.4 The Proposed Equating Method

After the estimation of the model proposed in the previous subsection, we obtain L samples from the posterior distribution of the cdfs of $X^\star$ and $Y^\star$. These estimations allow to obtain samples from the equipercentile function

$$\{\varphi_{Y^\star}^{(l)}(\cdot) = F_{Y^\star}^{-1(l)}(F_{X^\star}^{(l)}(\cdot)), l = 1, \ldots, L\}\ . \tag{12}$$

Given that each equipercentile function is computed from continuous cdfs, let us denote by $x_0^s, \ldots, x_{N_X}^s$ the original scores of test X rescaled into the support of the latent variable $X^\star$. Each of these rescaled scores are evaluated on $\varphi_{Y^\star}^{(l)}(\cdot)$ for $l = 1, \ldots, L$, thus obtaining L continuous equated scores for each x_k^s, i.e.

$$Z_{Y^\star, x_k^s}^\star = \{\varphi_{Y^\star}^{(l)}(x_k^s), l = 1, \ldots, L\}\ . \tag{13}$$

Finally, the equated discrete score for the score x_k is the score y_j, for some $j \in \{0, \dots, N_Y\}$, associated to the range $(\varrho_j; \varrho_{j+1}]$ (by means of Eq. (2)) that has the highest probability on the distribution of values (13). Thus, if $\varphi_Y(x_k)$ is the discrete equated score of x_k in the scale $\mathscr{Y}$, then:

$$\varphi_Y(x_k) = y_j \Leftrightarrow j = \max_{i \in \{0,\dots,N_Y\}} P\left(Z^{\star}_{Y^{\star},x_k^s} \in (\varrho_i; \varrho_{i+1}]\right) \tag{14}$$

Note that this model guarantees the symmetry property of equating functions because the equipercentile function is applied using continuous distribution functions.

3 Simulation Study

To illustrate the performance of the proposed equating method, we carried out a simulation study. In this section we describe how discrete test scores were simulated and how true discrete equated scores were obtained. Several simulated dataset were used in the simulation study. On each dataset we applied the proposed latent equating method and two traditional equating methods (Gaussian kernel equating and Equipercentile equating). Because these latter methods provide continuous equated scores, we define a method to make results comparable with discrete scores obtained from our latent equating method.

Under an equivalent group design, we considered tests with twenty items such that $N_X = N_Y = N = 20$. Scores X and Y were simulated considering the latent representation for ordinal variables described in Sect. 2.1. A mixture of two normal distributions was assumed for both latent variables Z_X and Z_Y. Discrete scores were obtained using the relation (2) where thresholds values were fixed to equidistant values between $\varrho_0 = -10$ and $\varrho_{N+1} = 10$. There is no loss of generality by considering the same thresholds values for both score test X and Y. This asseveration is based on the fact that both tests have the same number of items, and, because the fitting of the DPM model proposed for the latent variables in the model is independent of the thresholds values (Kottas, Muller, & Quintana, 2005).

Keeping in mind that latent variables considered in the simulation process were continuous random variables, we have a "real" version of the equipercentile function $\varphi_{Z_Y}(z_x)$. Using this function, true discrete equated values were obtained as the result of $\varphi_{Z_Y}(\gamma_i)$ where γ_i is the midpoint of the interval $(\varrho_i; \varrho_{i+1}]$ for $i = 0, \dots, N$.

Results shown in this paper are based on datasets simulated under the previous structure. Different sample sizes $n = (n_X, n_Y)$ were evaluated with $n_1 = (80, 100)$, $n_2 = (500, 500)$ and $n_3 = (1500, 1450)$. We simulated 100 replicates for each sample size. For each replicate, three Markov chains were generated, starting from different initial values. After completing a total number of 60000 iterations and a burn-in period of 30000 iterations, each chain was subsampled every 15 iterations. Combining these chains resulted in one chain of length 6000. The convergence of the chains was analyzed by computing the $\hat{R}$ statistic (Brooks & Gelman, 1998; Gelman & Rubin, 1992) which assess the between- and within-sequence variances of the chains. In addition, we also consider the effective sample size (Kass, Carlin, Gelman, & Neal, 1998). Results, not shown here, suggested convergence of the chains.

Model performance is summarized by computing the expectation, with respect to the sampling distribution, of the L_2-norm distance between the vector of true discrete equated scores (W) and the vector of estimated discrete scores ($\hat{W}$). This quantity was approximated using the Monte Carlo method and the 100 replicates generated for each sample size,

$$\Psi_2 = \mathbb{E}[\|W - \hat{W}\|_2] \approx \frac{1}{100}\sum_{i=1}^{100} \|W - \hat{W}_{(i)}\|_2 \,, \tag{15}$$

where $\hat{W}_{(i)}$ is the vector of estimated discrete scores $\hat{W}$ at the $i-$th replicate.

To compare the proposed method with current equating methods, the statistics Ψ_2 was also computed for equipercentile equating (EQ) and Gaussian kernel equating (KE). As already pointed out, because these methods produce equated scores that are actually continuous (i.e., not defined in the original discrete scale), in the evaluation of Ψ_2 we consider a discrete version of them using the largest integer number not greater than the corresponding continuous equated score. As a consequence, the discrete versions of traditional equated values can be properly compared with those obtained under the latent equating method.

The statistic Ψ_2 summarizes information over the whole score scale of the tests. To further analyze the performance of the proposed method, we also consider to evaluate the LE method among each possible value of the scale score. In order to do that, we compare true discrete equated scores and those obtained by the LE approach on each value of the scale score. Let us consider x_j a possible scale score on test X, for $j = 0, \ldots, N_X$. We computed the expected value of the difference between the true discrete equated score associated to the score x_j an its estimated discrete equated score under the proposed method. This expectation was approximated by using the 100 replicates for each sample size. To compare the proposed method with traditional equating methods, we also evaluate this quantity using both equipercentile equating and Gaussian kernel equating. Discrete equated scores estimated from traditional equating methods were obtained as the largest integer number not greater than the corresponding continuous equated score for each method.

3.1 Results

We summarize the results of applying the latent equating method on simulated data as described at the begining of this section. The evaluation of the proposed method is made over the whole scale score of the test as well as on each possible value of the scale score.

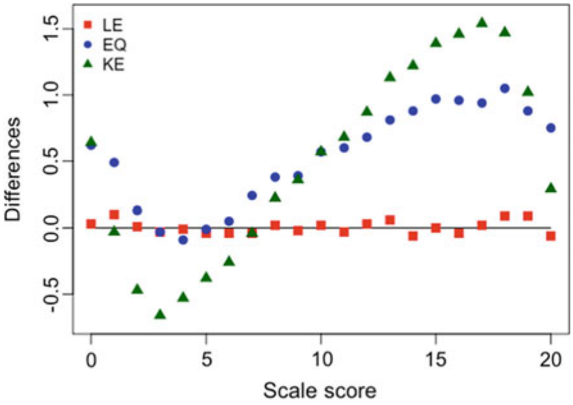
Values of the statistic Ψ_2 are summarized in Table 1 where it can be found the results of the latent equating method (LE), the equipercentile equating method (EQ) and the Gaussian kernel equating method (KE). On average, better estimations of the discrete equated scores were obtained using the proposed method. It can be seen that, for all samples sizes, our method outperforms the traditional equating methods. Note that these results are a summary of the performance of the proposed method and traditional equating methods on the estimation of discrete equated scores in the whole scale score.

To obtain a more detailed analysis of the proposed method's performance, we evaluated discrete equated scores estimated by the model on each possible score value on the scale. The same evaluation was made for the traditional equating methods EQ and KE but considering a discrete version of equated scores obtained under these methods. Results of this evaluation are summarized in Fig. 1. It can be seen that for almost all values on the score scale, equated values obtained using our method are much closer to the true equated values than those obtained from traditional equating methods.

Table 1 L_2-norm distance between the vector of true discrete equated scores and the vector of estimated discrete scores among three equating methods

Sample size	n_1			n_2			n_3		
Method	LE	EQ	KE	LE	EQ	KE	LE	EQ	KE
Ψ_2	5.03	6.67	5.98	2.95	4.78	4.72	2.05	3.97	5.20

Fig. 1 Mean differences between true equated values and its estimations on each possible scale score of test X



4 Conclusions and Discussion

Different parametric, semiparametric and nonparametric models have been proposed to estimate the equating transformation (González & von Davier, 2013). In all these methods, the equating transformation gives as a result continuous equated scores disregarding the fact that scores are actually defined on a discrete scale. In this paper we introduced an equating method that produces equated scores that are properly defined on the original discrete scale. Specifically, we develop a nonparametric Bayesian models for the score distributions through the use of a latent representation of ordinal variables. Results based on a simulation study have shown that, in comparison with discrete versions of equated values obtained by traditional equating methods, our approach has better performance in both the whole range of the scale and on each possible test score.

Although other approaches based on a Bayesian nonparametric model have been proposed (Karabatsos & Walker, 2009; González, Barrientos, & Quintana, 2015), we take advantage of the idea of Kottas et al. (2005) to obtain equated scores that are defined in the original scale of the tests. This idea, up to the best of our knowledge, has not been developed before.

The proposed approach can be extended in different ways. The DPM model can be replaced by alternative models that lead to estimate continuous probability distributions, such as Polya trees processes (Mauldin, Sudderth & Williams, 1992; Lavine, 1992) and mixture of Polya trees (Hanson & Johnson, 2002). Extensions of the proposed model could consider covariate-dependent Bayesian nonparametric models for the latent variables (MacEachern, 1999, 2000; De Iorio, Müller, Rosner, & MacEachern, 2004).

The proposed equating method was developed for samples from an equivalent group design. Extending the approach to other equating designs is a topic for future research. Also, applications to real data are planned for future work.

Acknowledgements Inés Varas was funded by CONICYT Doctorado Nacional grant 21151486. Jorge González was funded by FONDECYT grant 1150233. Fernando A. Quintana was partially funded by FONDECYT grant 1180034. This work was supported by Iniciativa Científica Milenio - Minecon Núcleo Milenio MiDaS.

References

Aitchison, J., & Silvey, S. (1957). The generalization of probit analysis to the case of multiple responses. *Biometrika*, *44*, 131–140.

Blackwell, D., & MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, *1*(2), 353–355.

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

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455.

De Iorio, M., Müller, P., Rosner, G., & MacEachern, S. (2004). An ANOVA model for dependent random measures. *Journal of American Statistical Association*, *99*, 205–215.

Ferguson, T. (1983). Bayesian density estimation by mixtures of normal distributions. In D. Siegmund, J. Rustage, & G. G. Rizvi (Eds), *Recent advances in statistics: Papers in honor of Herman Chernoff on His sixtieth birthday* (pp. 287–302). Bibliohound.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*, 209–230.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, *7*, 457–511.

Ghosh, S., Burns, C. B., Prager, D. L., Zhang, L., & Hui, G. (2018). On nonparametric estimation of the latent distribution for ordinal data. *Computational Statistics and Data Analysis*, *119*, 86–98.

Ghosh, J., & Ramamoorthi, R. (2003). *Bayesian Nonparametrics*. New York: Springer.

González, J., Barrientos, A. F., & Quintana, F. A. (2015). Bayesian nonparametric estimation of test equating functions with covariates. *Computational Statistics & Data Analysis*, *89*, 222–244.

González, J., & von Davier, M. (2013). Statistical models and inference for the true equating transformation in the context of local equating. *Journal of Educational Measurement*, *50*(3), 315–320.

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

Hanson, T., & Johnson, W. (2002). Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association*, *97*(460), 1020–1033.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*(453), 161–173.

Karabatsos, G., & Walker, S. (2009). A Bayesian nonparametric approach to test equating. *Psychometrika*, *74*(2), 211–232.

Kass, R. E., Carlin, B. P., Gelman, A., & Neal, R. M. (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *The American Statistician*, *52*(2), 93–100.

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

Kottas, A., Muller, P., & Quintana, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data. *Journal of Computational and Graphics Statistics*, *14*(3), 610–625.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, *20*(3), 1222–1235.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates I: Density estimates. *The Annals of Statistics*, *12*, 351–357.

MacEachern, S. (1999). Dependent nonparametric processes. In *ASA proceedings of the section on bayesian statistical science* (pp. 50–55).

MacEachern, S. (2000). *Dependent Dirichlet processes* (Tech. Rep.). Department of Statistics, The Ohio State University.

Mauldin, R., Sudderth, W., & Williams, S. (1992). Polya trees and random distributions. *The Annals of Statistics*, *20*(3), 1203–1221.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, *42*(2), 109–142.

McKelvey, R., & Zavoina, W. (1975). A statistical model for the analysis of ordinal dependent variables. *Journal of Mathematical Sociology*, *4*, 103–120.

Pearson, K. (1901). VII. On the inheritance of characters not capable of exact quantitative measurement. *Philosophical Transactions of the Royal Society A*, *195*, 79–150.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, *4*, 639–650.

Shah, D. A., & Madden, L. V. (2004). Nonparametric analysis of ordinal data in designed factorial experiments. *Phytopathology*, *91*(1), 33–43.

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

Comparison of Two Item Preknowledge Detection Approaches Using Response Time

Chunyan Liu

Abstract Response time (*RT*) has been demonstrated to be effective in identifying compromised items and test takers with item preknowledge. This study compared the performance of the effective response time (*ERT*) approach and the residual based on the lognormal response time model (*RES*) approach in detecting the examinees with item preknowledge using item response time in a linear test. Three factors were considered in this study: the percentage of examinees with item preknowledge, the percentage of breached items, and the percent decrease of response time of the breached items. The results suggest that the *RES* approach not only controls the Type I error rate below 0.05 for all investigated conditions, but also flag the examinees with item preknowledge sensitively.

Keywords Response time · Compromised items · Item preknowledge

1 Introduction

Item information on an examination can be divulged to test takers in different ways before they take the test. For example, previous test takers can share the items in person with future test takers or post the items on the Internet. On the other hand, examinees are motivated to obtain item information before the test because they tend to not only perform better on those items for which they have preknowledge, but they also tend to respond faster to those items, which will allow them more time to answer the unexposed items. Therefore, in order to ensure the integrity and validity of the test and to increase test security, it is necessary for test developers to accurately flag breached items and identify examinees with item preknowledge, especially for high-stakes examinations.

During computer-based testing (CBT) or computerized adaptive testing (CAT), both the examinee responses and response times (*RT*) are recorded. With the advancement of the statistical and measurement models of *RT*, more studies focus on *RT* as

C. Liu (✉)
National Board of Medical Examiners®, Philadelphia, PA 19104, USA
e-mail: cliu@nbme.org

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_31

an additional source of information in determining test speededness (Shao, Li, & Cheng, 2016), examinee motivation (Wise, 2006; Wise & Kong, 2005), and detecting aberrant examinee behavior (Meijer & Sotaridona, 2006; van der Linden & Guo, 2008; Qian, Staniewska, Reckase, & Woo, 2016). A correct response with an unexpected short amount of time may suggest that the examinee might have seen the item before taking the test.

Based on the loglinear model (van der Linden & van Krimpen-Stoop, 2003), Meijer and Sotaridona (2006) proposed the effective response time (*ERT*) approach to identify item preknowledge in CAT, in which the *RTs* for the able examinees' correctly answered items were used to predict the expected time needed for each item based on the examinees' ability and slowness. The difference between the observed and expected *RT* on the log scale was then used to flag examinees with aberrant behavior. Through a simulation study, Meijer and Sotaridona (2006) concluded that the higher the proportion of breached items and the greater the reduction of *RT*, the higher the power of detecting examinees with item preknowledge.

A lognormal *RT* model was proposed by van der Linden (2006), where the log scale of *RT* spent on each item were modeled as normally distributed. This model defines the discrimination and intensity parameters of *RT* for each item, and slowness parameter for each examinee. The residual of the observed and expected *RT* on the log scale from the model can be used to detect aberrant behavior. Qian et al. (2016) applied the lognormal *RT* approach on one non-adaptive exam and one adaptive exam. The residual of the log *RT* was used to detect the possible breached items and examinees with item preknowledge, and two items were flagged as breached items and two examinees were identified as potentially having item preknowledge. In this study, this approach is referred to as the residual (*RES*) approach. Details of the *ERT* and *RES* approaches are provided below.

1.1 *Effective Response Time (*ERT*) Approach*

Meijer and Sotaridona (2006) proposed the *ERT* approach to detect aberrant test behavior based on the loglinear item *RT* model (van der Linden & van Krimpen-Stoop, 2003), in which the log *RT* was defined as

$$lnt_{ij} = \mu + \delta_i + \tau_j + \varepsilon_{ij}, \tag{1}$$

with $\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$, where μ is the grand mean of log *RT* for the population of examinees on all items, δ_i is the *RT* parameter for item *i*, and τ_j is the slowness of the examinee *j*. Therefore, lnt_{ij} is normally distributed with a mean of $(\mu + \delta_i + \tau_j)$ and a standard deviation of σ. These parameters can be estimated as follows:

$$\mu \equiv E_{ij}\left(lnt_{ij}\right) \tag{2}$$

$$\delta_i \equiv E_j\left(lnt_{ij}\right) - \mu \tag{3}$$

$$\tau_j \equiv E_i\left(lnt_{ij}\right) - \mu \tag{4}$$

Meijer and Sotaridona (2006) defined the *ERT* of an item as the time required by an able examinee to answer the item correctly. In other words, the item *RTs* for those less able examinees and those who answer the item incorrectly will not be used in establishing the *ERT* for each item. The able examinees are selected such that the probability of getting a specific item right (P_{ij}) is larger than a prespecified cut value (γ), where P_{ij} is estimated based on the examinee's proficiency (θ) and the known item parameters for a given item response theory (*IRT*) model. It needs to be pointed out that the data used for estimating the *ERT* will be different for different items because the able examinee group will be different for different items.

The effective *RT* for each item *i* is modeled as

$$lnt_i = \beta_0 + \beta_1\boldsymbol{\theta} + \beta_2\boldsymbol{\tau} + \varepsilon, \tag{5}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ are vectors of examinee proficiency and examinee slowness, respectively, and considered as the known values that have been estimated previously ($\boldsymbol{\tau}$ is estimated using equation [4]), β_0, β_1, and β_2 are the regression coefficients, and ε is the error, which is considered as normally distributed. Therefore, the expected *RT* on the log scale can be estimated as the following:

$$\widehat{lnt}_i = \beta_0 + \beta_1\boldsymbol{\theta} + \beta_2\boldsymbol{\tau}. \tag{6}$$

For a specific examinee, *j*, the standardized difference between the observed *RT* and the expected *RT* on log scale for item *i* is considered to follow a standard normal distribution, that is,

$$z_{ij} = \frac{lnt_{ij} - \widehat{lnt}_{ij}}{\sigma_i} \tag{7}$$

where

$$\sigma_i = \sqrt{(J_i - 1)^{-1}\sum\nolimits_{j=1}^{J_i}\left(lnt_{ij} - \widehat{lnt}_{ij}\right)^2} \tag{8}$$

is the standard deviation of the *RT* on log scale for item *i,* and J_i is the number of examinees who took item *i*. Therefore, z_{ij}^2 follows a chi-square distribution with one degree of freedom, and the sum of the z_{ij}^2 over all the items will follow a chi-square distribution with the degree of freedom equal to the number of items (I), or, for each examinee *j*,

$$X_j = \sum\nolimits_{i=1}^{I} z_{ij}^2 \sim \chi^2(I). \tag{9}$$

The quantity of $\Pr\left(X_j \geq x\right) = p$ can be used to flag examinees with item preknowledge. If p is less than a prespecified significance level (eg, $\alpha = 0.05$), the examinee will be flagged as a possible candidate with item preknowledge.

1.2 Residual of Lognormal RT *Approach*

In the lognormal *RT* model (van der Linden, 2006), the log of the *RT* is considered to be normally distributed with a mean of $\left(\beta_i - \tau_j\right)$, which can be written as the following:

$$f\left(t_{ij}; \tau_j, \alpha_i, \beta_i\right) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left[\alpha_i\left(lnt_{ij} - \left(\beta_i - \tau_j\right)\right)\right]\right\}, \tag{10}$$

where τ_j represents the slowness of examinee j, α_i represents the discrimination parameter of item i, and β_i is the time intensity of item i (similar to the δ_i in the effective *RT* approach). For an examinee, the smaller the τ_j, the longer it will take him/her to respond to the item, and vice versa. These parameters can be estimated using the *MCMC* method (van der Linden, 2006).

The residual of the log *RT* is approximately a standard normal distribution, or $e_{ij} = \alpha_i\left(lnt_{ij} - \left(\beta_i - \tau_j\right)\right) \sim N(0, 1)$, which can be used to detect aberrant examinee behavior (Qian et al., 2016). In general, a correct item response with a large negative residue of the log *RT* may indicate that the examinee has preknowledge of this item.

Similar to the analogy in the *ERT* approach, the aberrant examinee behavior is detected using the chi-square statistic with a significance level of 0.05 since the sum of e_{ij}^2 over all the items follows a chi-square distribution with the degree of freedom equal to the number of items (I), or $X_j = \sum_{i=1}^{I} e_{ij}^2 \sim \chi^2(I)$. If the probability of $\Pr\left(X_j \geq x\right) = p$ is less than 0.05, the examinee is flagged as having item preknowledge. This approach is referred to as the residual (*RES*) approach in this study.

Both the *ERT* approach and the *RES* approach have been investigated to detect item preknowledge. However, no studies have been conducted to compare these two approaches in detecting item preknowledge. It is worth comparing the performance of these two approaches in terms of Type I error and power of detecting examinees with item preknowledge. More specifically, the following factors will be considered: the percentage of test takers with item preknowledge, the percentage of breached items, and the percent decrease of the *RT* of the breached items.

2 Data

The data used in this study are from a high-stakes medical licensure examination. The test, administered in 2017, includes about 280 multiple-choice items, which are

Table 1 Summary of the data

	Raw Score	b	α	β	Reliability
Mean	24.45	−0.01	1.44	4.29	0.76
SD	4.55	0.74	0.71	0.31	

divided into seven 60-minute blocks. In the current study, only a subset of items (32 items) from one block were used ($N = 6611$). It needs to be noted that, considering the test design, test administrations, form spiraling, and the comparison of examinee performance with previous years, no clear evidence indicates examinees with item preknowledge or collusion in this sample. In addition, given the fact that the test takers have spent several years in medical school and are required to pass the examination before being permitted to practice medicine, it is very unlikely that the test takers have low motivation or rush to the end of the test by providing rapid guessing.

The items were calibrated using a 1-PL logistic *IRT* model. These item parameter estimates were considered as known in this study. Table 1 provides the summary statistics of the raw score, *IRT* difficulty estimates, α (discrimination parameter of item response time) and β (time intensity) of item from the lognormal model, and Cronbach's alpha.

3 Methods

In this study, the *ERT* and *RES* approaches were compared in two situations in terms of the item preknowledge detection: without item breach and with item breach. In the situations with item breach, three factors were considered: (1) percentage of examinees with item preknowledge (25, 50, and 75%), (2) percentage of breached items (10, 25, and 50%), and (3) percent decrease of *RT* of the breached items (25, 50, 67, and 75%). In the situations with item breach, it needs to be pointed out that only the *RTs* were manipulated because the purpose of this paper is to compare the two approaches using *RT* only.

3.1 Situation Without Item Breach

For the situation without item breach, one thousand test takers were randomly drawn without replacement from the original data. Both the *ERT* and *RES* approaches were implemented to the sampled data to flag examinees with item preknowledge using a significance level of 0.05. This process was repeated 100 times and the Type I error is estimated as the average proportion of the flagged examinees.

3.2 Situations With Item Breach

In this situation, the following steps were applied for each of the manipulated conditions:

1. Randomly draw 1000 test takers without replacement from the original data.
2. Randomly sample examinees with item preknowledge without replacement from the sample obtained in Step 1.
3. Randomly sample items that are breached without replacement.
4. Reduce the *RT* of the selected breached items for those examinees with preknowledge based on the percent decrease of *RT* for a given condition.
5. Flag examinees with item preknowledge using both *ERT* and *RES* approaches.
6. Calculate the Type I error and power in flagging the examinees with item preknowledge.
7. Repeat Steps one to six 100 times and calculate the average Type I error and power.

In this study, the item and person parameters for the *RES* approach were estimated using the *MCMC* sampling method with 6000 iterations through the *LNIRT* R package (Fox, Entink, & Klotzke, 2017). For the *ERT* approach, the pre-specified cut value (γ) was set to be 0.4.

4 Results

In this section, the results of the log *RT* model fit are presented first, followed by the results of the Type I error of the two approaches for the original data without item breach. The comparison of the Type I error and power at different manipulated conditions are provided last.

4.1 Model Fit of Log RT

Because both approaches assume that the log *RT* is normally distributed, the *RT* was plotted first to examine whether the data can be fitted using the lognormal model before evaluating the performance of the two approaches in detecting item preknowledge. The fit of the lognormal model to *RT* was evaluated by examining (1) the fit of the lognormal model to the frequency distribution of *RT* data, and (2) the relationship of the observed and predicted cumulative distribution of the *RT*. If the model fits the data perfectly, the curve of the observed versus predicted cumulative distribution of *RT* should be identical to the identity line (van der Linden, 2006; van der Linden & Guo, 2008).

Figure 1 provides the visualization of the distribution of RT for a typical item (item 10) and the examination of the model fit using lognormal model (red dashed

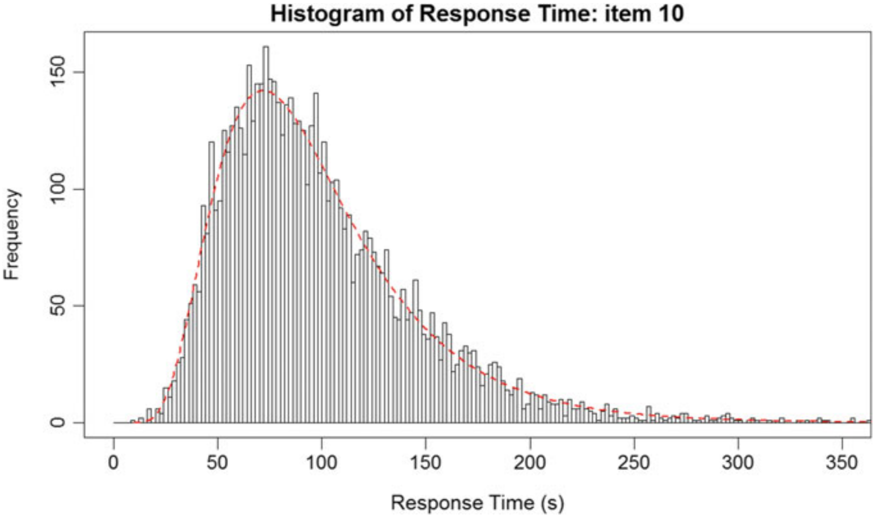


Fig. 1 Examination of lognormal model fit of item *RT*

line). The histogram of RT indicates that the distribution of RT is right skewed and lognormal model can fit the data quite well for this item. Although not provided, the plot of the observed versus predicted cumulative distribution of the log RT suggests that the lognormal model fits this item almost perfectly because the observed cumulative distribution and the predicted cumulative distribution are almost identical. In addition, the Kolmogorov–Smirnov goodness-of-fit Test was also used to examine model fit for all items. The results suggested that out of the 32 items, the response times for 28 items can fit the lognormal distribution at a 0.05 significance level. Overall, it is concluded that the lognormal model fits the log RT reasonably well for all items, especially consider the large sample size.

4.2 Situation Without Item Breach

Using the original data without manipulation of the *RT*, the average Type I error was 0.130 and 0.014 for the *ERT* and the *RES* approaches, respectively, when the nominal significance level was 0.05. Therefore, it is concluded that the *ERT* approach cannot control the Type I error rate and the *RES* approach is slightly conservative when no item is breached. However, these results may be caused by the potential existence of examinees with item preknowledge or rapid guessing in this real dataset, even though there is no clear evidence of item breach or low examinee motivation.

4.3 Situations With Item Breach

Under each of the conditions described in the Methods section, Type I error and power of flagging examinees with item preknowledge were estimated for both *ERT* and *RES* approaches. Figure 2 provides the comparison of Type I error rate (upper) and power (lower) when 50% of the examinees have item preknowledge to some items. The plots for other conditions are not presented but described below due to the page limit.

Figure 2 suggests that when 50% of examinees have item preknowledge, the Type I error is about 0.13 for the *ERT* approach and from 0.014 to 0.032 for the *RES* approach. Although the Type I error tends to increase slightly with the increase of the percentage of breached items and percent decrease of response time for both *ERT* and *RES* approaches, the effect is small and almost negligible. The plot of power (bottom) indicates that the *RES* approach is very sensitive to both the percentage of breached items and percent decrease of response time. As the percent decrease of response time increases, the power increases significantly. For example, the power is about 0.02, 0.10, 0.27, and 0.72 when the percent decrease of *RT* is 25, 50, 67, and 75% and the percentage of breached items is 25% (red line). Similarly, the power increases as well when the percentage of breached items increases. These results suggest that, for the *RES* approach, the more breached items and the more reduction of item *RT* of the breached items, the more likely the examinees with item preknowledge will be detected. However, the *ERT* approach is not sensitive in flagging examinees with item preknowledge.

Although not presented in this paper, the results suggest that when 50% of the items were breached, the Type I error for the ERT approach is larger than 0.12 for all conditions, but the RES approach controls Type I error very well (<0.05). Type I error tends to increase slightly with the increase of percentage of examinees with item preknowledge and the percent decrease of RT for both methods. However, the effect is not significant. In addition, the power of the RES approach is sensitive to both the percentage of examinees with item preknowledge and percent decrease of RT. The power increases as the percent decrease of *RT* and the percentage of examinees with item preknowledge increase. More specifically, the power is about 0.03, 0.21, 0.65, and 0.99 when the percent decrease of *RT* is 25, 50, 67, 75 and 50% of examinees with preknowledge. This result suggests that, for the *RES* approach, the more examinees with item preknowledge and the more reduction of item *RT* of the breached items, the more likely the examinees with item preknowledge will be detected. However, this was not observed for the *ERT* approach.

The conclusions can be summarized as the following when the percent decrease of *RT* is fixed at 67%: (1) the *RES* approach controls the Type I error very well, but not the *ERT* approach; (2) the Type I error tends to increase slightly with the increase of the percentage of examinees with item preknowledge and the increase of the percentage of breached items for both approaches; (3) for the *RES* approach, the power increases significantly with the increase of the percentage of breached items but slightly with the increase of the percentage of examinees with item preknowledge;

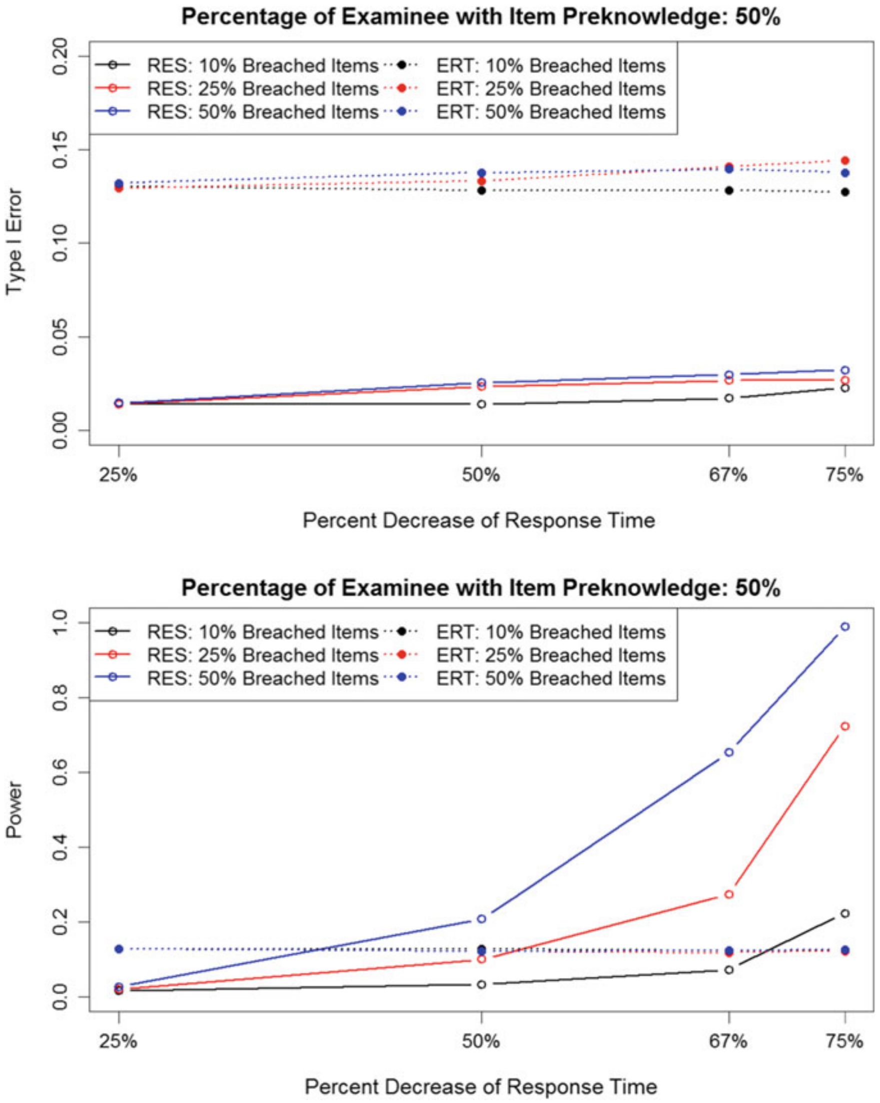


Fig. 2 Comparison of type I error (upper) and power (lower)

and (4) both the percentage of examinees with item preknowledge and the percentage of breached items have little effect on power for the *ERT* approach.

5 Conclusion and Discussion

This study compares the performance of the *ERT* approach and the *RES* approach in detecting the examinees with item preknowledge using item *RT* only for a linear test. The results suggest that the *RES* approach not only controls the Type I error below 0.05 for all investigated conditions, but also flags the examinees with item preknowledge sensitively. More specifically, the *RES* approach is more likely to identify examinees with item preknowledge when the sample contains a higher percentage of examinees with item preknowledge, more breached items, and a greater reduction of the *RT* for breached items.

In general, the power of identifying examinees with item preknowledge is the most sensitive to the reduction of *RT* when items are breached. In the situations that the percent decrease of *RT* of breached item is 50% or less than the original *RT*, the power of correct detection of examinees with preknowledge is very low (<0.25). However, the power can be as high as almost 1 when the percent decrease of *RT* is 75%. The *RES* approach is also sensitive to the percentage of breached items. If the number of breached items is less than 10%, the power is less than 0.2 for all simulated conditions. As the percentage of breached items increases, the power increases significantly. These conclusions make sense since the higher the reduction of *RT* and the more breached items, the larger the difference between the observed and expected log *RT* for the breached items and the larger the chi-square statistic. The percentage of examinees with item preknowledge has relatively less impact on the performance of the *RES* approach.

Meijer and Sotaridona (2006) concluded that the *ERT* approach was sensitive to identify item preknowledge in CAT through a simulation study. They found that the Type I error was below 0.05 and that the detection rate was about 0.95 and 0.5 in the conditions where the *RT* of the breached items was 25 and 50% of the original *RT*, respectively. However, this was not observed in the current study. This study suggests that the *ERT* approach is not effective in identifying examinees with item preknowledge since it can neither control the Type I error nor identify the examinees with item preknowledge accurately. This might be due to the differences between the Meijer and Sotaridona (2006) study and the current study. To summarize, the Meijer and Sotaridona (2006) study used a computerized adaptive test setting where items were administered to examinees with similar ability, and all examinees were considered as having item preknowledge in all simulated conditions. In addition, compared to the data used in this study, the *ERT* approach may fit the data much better in the Meijer and Sotaridona (2006) study.

The current study evaluated the performance of two approaches in identifying examinees with item preknowledge based on *RT* from real data only. The limitation and future direction are summarized as follows. First, item responses can also provide more information about whether an examinee has preknowledge since it is more likely that an examinee gets an item right if he/she was exposed to the item before the test. Therefore, the combination of *RT* and item responses should be considered for flagging examinees with preknowledge for future studies. Second, given the high-stakes of the examination used in this study, we assumed that all examinees took the exam with high motivation and didn't consider faster *RT*s due to rapid guessing or lost motivation. In the future study, the examinees associated with rapid guessing or low motivation will be excluded. Third, given the existence of misfit of the *RT* to the lognormal distribution and the potential examinees with rapid guessing, more studies should be focused on simulated data, instead of real data.

Test security has gained a lot of attention from testing organizations. However, it needs to be pointed out that even if an examinee is statistically flagged as having preknowledge for some items, action needs to be taken prudently in practice. The test organizations should consider the possible consequences, and possible legal and practical implications when notify the examinees of their cheating behaviors.

References

Fox, J. -P., Klein Entink, R. H., & Klotzke, K. (2017). LNIRT: LogNormal response time item response theory models. R package version 0.2.0. Retrieved from http://CRAN.R-project.org/package=LNIRT.

Meijer, R. R., & Sotaridona, L. S. (2006). Detection of advance item knowledge using response times in computer adaptive testing. (LSAC Computerized Testing Report No. 03-03). Newtown, PA: Law School Admission Council.

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice, 35,* 38–47.

Shao, C., Li, J., & Cheng, Y. (2016). Detection of test speededness using change-point analysis. *Psychometrika, 81,* 1118–1141.

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

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31,* 181–204.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73,* 365–384.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes, computer-based test. *Applied Measurement in Education, 19,* 93–112.

Wise, S. L., & Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18,* 163–183.

Identifying and Comparing Writing Process Patterns Using Keystroke Logs

Mo Zhang, Mengxiao Zhu, Paul Deane and Hongwen Guo

Abstract There is a growing literature on the use of process data in digitally delivered assessments. In this study, we analyzed students' essay writing processes using keystroke logs. Using four basic writing performance indicators, writers were grouped into four clusters, representing groups from fluent to struggling. The clusters differed significantly on the mean essay score, mean total time spent on task, and mean total number of words in the final submissions. Two of the four clusters were significantly different on the aforementioned three dimensions but not on typing skill. The higher scoring group even showed signs of less fluency than the lower scoring group, suggesting that task engagement and writing efforts might play an important role in generating better quality text. The four identified clusters further showed distinct sequential patterns over the course of the writing session on three process characteristics and, as well, differed on their editing behaviors during the writing process.

Keywords Writing process · Keystroke logs · CBAL · Sequential pattern · Editing behavior

1 Background

Writing is an important skill that is taught in schools and valued in the workplace. Writing tasks, such as the essay writing examined in this study, are frequently included in standardized assessments, licensure qualification tests, placement tests,

M. Zhang (✉) · M. Zhu · P. Deane · H. Guo
Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08540, USA
e-mail: mzhang@ets.org

M. Zhu
e-mail: mzhu@ets.org

P. Deane
e-mail: pdeane@ets.org

H. Guo
e-mail: hguo@ets.org

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_32

portfolio-based performance assessments, as well as in-class formative assessments. Normally, essays are graded by human raters and/or automated scoring systems based on a holistic scoring rubric. Despite the benefits and ubiquitous use, a holistic score contains rather limited information and offers scant, if any, instructional feedback that can be returned to the test users (e.g., students, teachers, district officials, admissions officers). In other words, holistic scores provide an overall evaluation of the writing quality, but do not pinpoint the specific areas where a writer might struggle. For digitally-delivered writing assessments, however, we can record the moment-by-moment *processes* by which writers generate their responses through keystroke logging. In contrast to a holistic score, much finer-grained information about students' writing proficiency can be obtained from analyzing the keystroke logs. In addition to the usual types of feedback generated by automated scoring systems, which are often based the final product (e.g., a written text, a spoken response), keystroke logs allow feedback given based on the item response process. From the keystroke logs, one can identify, for example, whether a writer has spent sufficient time and effort on the task, has trouble with spelling, encounters difficulties in typing on the keyboard, or has edited what was written before turning in the response.

In this study, we used a keystroke logging system developed at Educational Testing Service – one primarily intended for large-scale digital administrations to support classroom instruction and educational assessment (Deane et al., 2016; Zhang, Bennett, Deane & van Rijn, 2019). A well-designed keystroke logging system records all the changes to the text buffer while a student is writing, along with associated time stamps. Additional information that may be tracked include cursor movements, mouse clicks, and access to resources outside the text entry window, such as time spent reading external references and the use of editing tools supplied by the task-delivery interface. The entire text production process can be reconstructed from the keystroke log, although additional processing is required to extract meaningful information from the raw keystroke log. Depending on the construct to be measured, it may be necessary to identify linguistic features of the text using natural language processing techniques and define appropriate classifiers with which to characterize writer's performance patterns. These classifiers, often called "features," provide the input for higher-level analyses of writing patterns. See Zhang, Hao, Li, and Deane (2018) for an example of developing a "burst" measure of writing translation from keystroke logs.

There is a sizable body of research that has applied keystroke logging to writing research in general, but the research on the use of keystroke logs in the testing or educational assessment context has just started to emerge. Sinharay, Zhang, and Deane (2019) argue that feedback about students' writing processes, summarized at the group or individual levels, may help classroom teachers make their instructional decisions. Previous research in the assessment context has reported that a number of timing and process features, such as burst length, in-word typing speed, between-word pause length, and initial pause time before typing a word, are indicative of writing proficiency and worth reporting (Guo, Deane, van Rijn, Zhang, & Bennett, 2018; Zhang & Deane, 2015). Information about the writing processes revealed from keystroke logs enhances the feedback given to writers by going beyond simply "high" and "low" scores, and potentially rendering rich profiles of their writing process and proficiency on various dimensions (e.g., task planning, editing, fluency).

In this study, we addressed two research questions. The first research question (RQ1) is: *Can we identify students' writing patterns by applying clustering techniques to the most basic performance indicators (i.e., time spent on task, essay score, number of keystrokes, and typing speed)?* The second research question (RQ2) is: *How do students' writing processes and editing behaviors differ across clusters?* The answers to these questions will have implications for the classification of students' writing patterns and identifying possible weak areas in writing, particularly from the perspective of editing behaviors, which can be useful information for the instructors and writers/students for the improvement of teaching and learning of writing.

2 Methods

2.1 Participants and Instrument

We used a data set collected from a research program at Educational Testing Service, the Cognitively Based Assessment *of*, *for*, and *as* Learning (CBAL®). The CBAL summative writing assessment is a scenario-based assessment, in which the items are designed around a unifying scenario and are sequenced according to a theoretically determined order (Bennett, Deane, & van Rijn, 2016). In this study, we focus on the timed culminating essay writing task. Students were given 35 min to complete their essay, though they were allowed to submit sooner before the time ran out. The data set was collected as part of an experimental study involving a number of schools across the United States (van Rijn, Chen, & Yan-Koo, 2016; van Rijn and Yan-Koo, 2016). As an initial investigation, we used the data collected in one form, which we will refer to by the shorthand label "Culture Fair," since the assessment is built around a scenario in which students must choose the best theme for their school's Culture Fair event.

After excluding a small number of responses where the keystroke logs were corrupted due to system failure, and a small additional set of essay responses where students submitted blank, meaningless, or off-topic responses, our final data set included 740 qualified essay responses submitted by 6th- to 8th-grade students. There were 276 male students and 285 female students. About 24% did not report their gender. The majority of the students were English native speakers (62%). As for ethnicity, 52% were White, 17% Hispanic, 3% African American, 3% Asian, less than 1% belonged to other groups, and (the same) 24% did not report on this question. All responses were scored against two different rubrics: one focused on general writing quality and another focused on the quality of the content specific to the writing genre. On each rubric, essays were double-scored by a pair of randomly-assigned, trained raters, achieving an inter-rater reliability of 0.70 and 0.72 as measured by quadratically-weighted kappa for the two rubrics, respectively, as reported in van Rijn et al. (2016). For the purpose of analysis, we used the first human rating on each rubric, which is also a common practice. Finally, all the essay responses were parsed by the ETS keystroke logging engine.

2.2 Clustering of Writing Logs (RQ1)

To address the first research question, we used four variables to characterize writing performance: essay score, total number of keystrokes, keyboarding speed, and total time spent on task. These variables provide basic information about writing performance. Specifically, these variables are straightforward and provide information about related, but distinguishable, aspects of writing effort and fluency. Table 1 gives the inter-correlations between these variables in our data set. The *essay score* represents the overall quality of the final response. It is calculated as the sum of the two rubric scores, and is moderately related to the other variables. The *total number of keystrokes* is, at least to some degree, a proxy for the overall effort that a writer makes during the writing process. This count includes every action the writer makes, including insertions, deletions, cuts, pastes, and jumps. The *total time on task* is another indicator of writing effort. This variable is calculated as the total time elapsed between the first and last keystrokes. However, it should be noted that a longer time on task does not necessarily mean more keystrokes, or vice versa. The Pearson's correlation between the total number of keystrokes and total time on task, in our data set, is 0.72 ($p < 0.0001$) (Table 1). Finally, the *in-word keyboarding speed* represents the writer's familiarity with the keyboard. It is measured in terms of the characters produced per second while typing the most common English words (Zhang, Feng, Deane, & Guo, 2018). This measure is defined using only common words that were typed correctly without any editing, in order to obtain a purer measure of keyboarding skill. As a result, this speed is likely to be the upper-bound of a writer's typing speed. Typing speed, so defined, is only moderately related to essay score (Pearson's $r = 0.31$) and to the total number of keystrokes produced in the writing process (Pearson's $r = 0.40$), and, as expected, is weakly and negatively related to total time spent on task (Pearson's $r = -0.07$).

We clustered students on these four variables using Ward's minimum-variance method (Ward, 1963). The central notion in Ward's method is to minimize the within-

Table 1 Pearson's correlations between clustering input variables

Clustering variable	1	2	3
1. Essay score (2–10)			
2. Number of keystrokes	0.59		
3. Total time on task (in second)	0.45	0.72	
4. In-word keyboarding speed (in character/second)	0.31	0.40	−0.07

Note All correlations are significant at $p < 0.0001$, N = 740
Copyright by Educational Testing Service, 2019 All rights reserved

cluster variance using a distance algorithm based on the sum of squares. Murtagh and Legendre (2011, 2014) provide the algorithms and details of this method. Several indices were considered to determine the optimal number of clusters. The cubic clustering criterion (CCC) was computed, for which a local peak on CCC indicated a good clustering. We also used the Pseudo F statistic for which the locally highest value suggested the most desirable number of clusters. R-squared and semipartial R-squared were also considered in making the decision. For clustering analyses, it is helpful to visually examine the separation (or overlap) between the clusters' locations on a n-dimensional space, where n refers to the number of input variables. However, it is impossible to project a n-dimensional space if $n > 3$. Canonical discriminant analysis, as a dimension reduction method, offers one solution to this problem (Rencher, 1992). Canonical discriminant analysis finds the linear combination of the input variables that can explain the largest proportion of the between-cluster variance. Cooley and Lohnes (1971) gives the mathematical basis for the canonical discriminant analysis. In this study, we generated the scatter plot between the canonical variables and visually examined the degree of separation between identified clusters. The analyses were conducted using SAS®. Specifically, the function *proc cluster* was used with *method* = *ward* and specifications of *ccc*, *pseudo*, and *rsquare*, as evaluation criteria; and the function *proc candisc* was applied for the canonical discriminant analysis. The identified clusters were then interpreted based the distributions of the clustering variables. MANOVA and Tukey's post-hoc multiple comparisons were further conducted to compare the means of input variables between clusters. The dependent variables in MANOVA were the four clustering variables, and independent variable was the cluster assignment. Pillai's Trace and F statistics were used to evaluate the model effect.

2.3 Comparing the Clusters (RQ2)

Once clusters were identified, we compared the clusters based on two aspects. First, we compared the sequential patterns during writing across clusters. That is, we examined how the values of specific variables changed at different stages in the writing process. We examined three specific variables (or characteristics): the median pause interval between keystrokes (an indicator of fluency), the ratio of deletions to insertions (an indicator of editing behavior), and general writing speed (total number of keystrokes divided by time on task) as another indicator of fluency. These were the features analyzed in Zhu, Zhang, and Deane (2019) where the authors studied the sequential patterns on these features for different gender and ability groups. Statistical transformations were taken on two of the three features for the purpose of normalization (Table 2).

To capture the changes in one's writing process within a writing session, each keystroke log sequence for each student was evenly divided into 30 segments, based on the total active writing time (i.e., time between first and last keystrokes). This way, each log sequence was reduced to 30 data-sampling points, each corresponding to a

Table 2 Process variables used for analyses of sequential patterns

Short name	Calculation	Interpretation	Transformation
MedIKI	Median pause interval between keystrokes	Overall composition fluency	Median of Log(IKI)s
DIRatio	Ratio of deletion over insertion	Extent of editing of any kind	Log((del +1)/ins)
GenSpeed	Number of keystrokes per second	General writing speed	

Table 3 Editing-related summary features

Summary editing feature	Definition
AmountofDiscardedText	Percentage of deleted characters as a function of total number of keystrokes
RateTypoCorrection	Log ratio of number of corrected typographical errors over the number of uncorrected typographical errors
MedianJumpLength	Median distance of all jumps
ProportionPreJump-PauseTime	Percentage of time spent before making a jump as a function of total time on task
MedMaxWordEdit-PauseTime	Median length of the longest in-word pauses
ProportionEditedChunks	Percentage of edited sequences in terms of number of keystrokes produced as a function of total number of keystrokes
ProportionEditedWords	Percentage of words that were edited as a function of total number of words written during text production including those deleted later
ProportionMinorEdits	Percentage of text that were edited with no more than two characters as a function of total number of keystrokes
ProportionMultiWord-Delete	Percentage of deleted sequences that were edited with no more than two characters as a function of total number of keystrokes

subsession in the writing process, which also allowed us to identify corresponding points in the writing process across keystroke logs. We aggregated the data for each cluster, and calculated the median value for each variable in Table 2 at each of the 30 sampling points for members of that cluster.

We ran a simple linear regression for each cluster by regressing the characteristic of interest on the subsession sequence (1–30) and compared the intercepts and slopes between the clusters. In cases where the trend was not linear, we also applied a non-parametric approach – LOESS smoothing – to compare and visualize the trends on each characteristic between clusters (Cleveland and William, 1979). The 95% confidence interval bands were also generated. The SAS function *proc loess* with *smooth = 0.6* was used.

In addition to the sequential patterns over time, we compared the clusters, based on a number of selected summary features that were designed to characterize editing behaviors. The summary features of interest are described in Table 3. The summary features were computed based on the entire keystroke log (as opposed to by subsession). We ran MANOVA with the independent variable being the cluster assignment and dependent variables being the summary editing features. Tukey's post-hoc multiple comparisons were conducted to detect any significant mean differences between clusters. This analysis would further assist us in explaining the differences between clusters in a more substantive way from the perspective of text-editing behaviors.

3 Results

3.1 Clustering of Writing Logs (RQ1)

Based on the Ward's minimum variance cluster analysis, all clustering criteria suggested four clusters. Both the CCC and pseudo F statistics peaked at four clusters (Fig. 1), while the four clusters explained 80% of the total variance. The elbow turning points for the R-squared and semipartial R-squared were also located at the four clusters (not shown), suggesting that the increase in R-squared became small once going beyond four clusters. Based on these results, we decided to move forward with four clusters as the most parsimonious solution.

Canonical discriminant analysis was used to visually examine the separation between the clusters. The result of the canonical discriminant analysis revealed that, in our data, the first two canonical variables explained nearly 100% of the between-cluster variance (Table 4). Hence it is reasonable to project the four clusters on a two-dimensional plane. The scatter plot between the first and second canonical variables (Fig. 2) shows that the four clusters are rather separated from one another, which further justifies the four cluster solution.

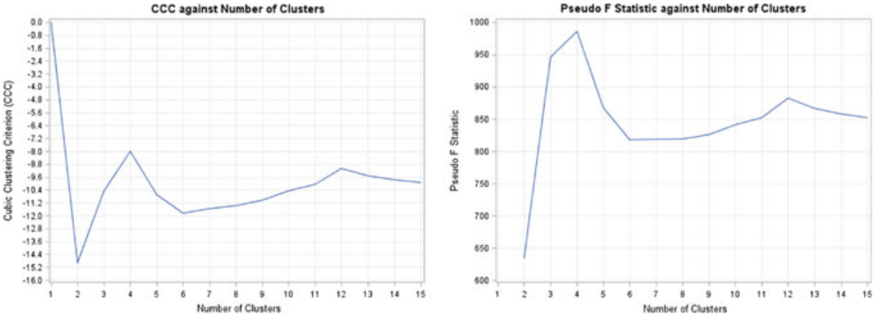


Fig. 1 Clustering criteria results. *Note* CCC (left) and Pseudo F (right) statistics against the number of clusters

Table 4 Variance explained by canonical variables

Canonical variable	Canonical correlation	Squared canonical correlation
1	0.94	0.89
2	0.41	0.17
3	0.10	0.10

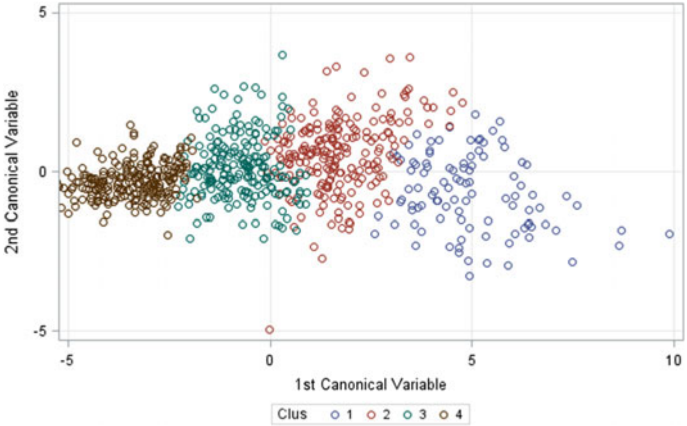


Fig. 2 Two-dimensional plot between first and second canonical variables

To interpret the clusters, Table 5 gives the means of each clustering variable for each cluster. MANOVA and post-hoc multiple comparisons were further carried out to compare the means. MANOVA results indicated a statistically significant overall effect for cluster, with the MANOVA test statistic Pillai's Trace = 1.07, F(12, 2205) = 101.79, $p < 0.0001$. For each column in Table 5, the same letter superscription (in a column) indicates that the means are not statistically different, whereas the different letter superscription indicates that the means are significantly different.

The results show that the number of students falling into each cluster is relatively even ranging from 103 to 217. The MANOVA analyses revealed that the four clusters were significantly different on their mean essay scores, mean total number of keystrokes produced during the text production, and mean total time spent on writing.

Cluster 1 has the highest mean essay score (6.73 on the scale of 2–10), spends on average the longest time on writing, produces the most number of keystrokes, and shows the fastest keyboard typing speed. These patterns indicate that Cluster 1 consists of relatively more fluent and capable writers. Writers in Cluster 2 receive, on average, lower essay scores than writers in Cluster 1; they also display a slower typing speed and spend, on average, 3.4 min less on writing than writers assigned to Cluster 1. It is noteworthy that Cluster 2 writers produce significantly fewer keystroke actions on average than Cluster 1 writers (1533 in Cluster 2 vs. 2426 in Cluster 1), possibly because they are much less fluent typists. Cluster 3 – the next lower scoring

Table 5 Results of clustering

Cluster	N	Essay score (2 to 10)	Number of keystrokes	Total time on task (in second)	Keyboarding speed (in character/second)
1	103	6.73[a]	2426[a]	1402[a]	5.25[a]
2	216	6.04[b]	1533[b]	1198[b]	4.28[b]
3	204	5.32[c]	1105[c]	804[c]	4.42[b]
4	217	4.13[d]	596[d]	429[d]	4.05[c]

Note A different superscription letter in each column suggests a significant mean difference at $p \leq 0.05$ within a column

group than Cluster 2 – spends, on average, 804 seconds (13.4 min) on writing, significantly less than Clusters 1 and 2. Compared to the highest scoring group Cluster 1, Cluster 3 writers produce less than half of the keystroke actions. Cluster 3 writers also produce fewer keystrokes than Cluster 2; however, their typing speed is *not* significantly different from Cluster 2 writers. It is possible that the low mean essay score for Cluster 3 writers is due to a lack of writing effort, low motivation, problematic writing strategies, and/or other linguistic struggles. Finally, Cluster 4 writers appear to be struggling with the writing task: they have the lowest mean human scores (4.13 on average), spend only on average 429 seconds (about 7 min) on the task, demonstrate the slowest average in-word typing speed among the four clusters, and produce only 596 keystrokes on average during the writing process.

In sum, findings in RQ1 show that, using the four basic writing performance indicators, students' logs can be grouped into four meaningful clusters, in which higher scores correlate with more time on task, faster keystrokes, and a larger total number of keystrokes produced. Additionally, writing efforts and motivation may play a role in generating higher quality text.

3.2 *Comparing Clusters (RQ2)*

Next, we compared the four clusters based on two sets of features that were not directly used for the clustering purpose. First, the clusters were compared based on sequential patterns on three writing process characteristics. Second, the clusters were compared based on a selected set of features that summarized different editing behaviors.

Comparing Sequential Patterns. The sequential pattern over the course of writing process for each cluster was examined based on the estimated parameters of intercept and slope using simple linear regression; that is, regressing the process characteristics on writing subsessions 1–30. The trends were also visually examined using non-parametric LOESS smoothing for each of the three characteristics. The three characteristics were described earlier in Table 2: the median pause interval

Table 6 Estimated parameters of simple linear regression

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
	(n = 103)	(n = 216)	(n = 204)	(n = 217)
MedIKI				
Intercept	−1.386 (0.010)***	−1.197 (0.008)***	−1.234 (0.008)***	−1.167 (0.014)***
Slope	−0.001 (0.001)	−0.001 (0.000)*	−0.000 (0.000)*	−0.000 (0.000)
DIRatio				
Intercept	−1.918 (0.033)***	−2.084 (0.030)***	−2.226 (0.039)***	−2.547 (0.047)***
Slope	−0.006 (0.002)*	−0.009 (0.002)***	−0.005 (0.002)*	0.002 (0.003)
GenSpeed				
Intercept	1.724 (0.034)***	1.326 (0.036)***	1.585 (0.048)***	1.682 (0.055)***
Slope	0.005 (0.002)*	0.002 (0.002)	0.000 (0.003)	−0.001 (0.003)

Note ***Statistically significant at $p < 0.0001$, **$p < 0.001$, *$p \leq 0.05$. Standard errors are in parentheses

between keystrokes (logged), ratio of deletion over insertion (logged), and number of keystrokes per second.

Table 6 presents the intercept and slope parameter estimates of the simple linear regression. The regressions were run separately for each cluster. An assumption of simple linear regression is that the relationship between time (writing subsessions 1–30) and writing characteristic (i.e., MedIKI, DIRatio, and GenSpeed) is linear. The patterns of their relationships can be directly observed, however, in the corresponding plots (Fig. 3), which suggest a nonlinear trend in some cases.

For the MedIKI, the result indicates that Cluster 1 has much shorter median IKIs throughout the writing process than the other clusters. The first panel in Fig. 3 shows that, for Cluster 1 (first left), the MedIKIs are slightly higher at the beginning of writing, suggesting slower writing at the start of writing process. And the level of fluency, as indicated by logged median pause intervals between keystrokes, is increased in the middle and slightly dropped again towards the end of the writing process. The result of the insignificant estimated slope parameter of the simple linear regression (−0.001) is possibly due to this potential quadratic pattern. The quadratic pattern between time and MedIKIs observed in Cluster 1 is not observed for Cluster 2. For Cluster 2, there appears to be an increasing trend on fluency (indicated by MedIKI) as writing proceeds, based on the significant slope parameter estimate, although the degree of the slope (−0.001) is rather minimal. It is fair to conclude that, for Cluster 2, the level of fluency appears to be quite consistent over the course of text production. Cluster 3 shows a similar pattern to Cluster 2 with a consistent level of fluency over the course of the writing process. The level of fluency in Cluster 3 even appears to be slightly faster than Cluster 2. Cluster 4, in general, shows the lowest level of fluency as measured by MedIKI among all clusters. As for the sequential pattern, Cluster 4 does not exhibit either a linear or a quadratic trend. From the far-right figure in the top panel in Fig. 3, a quarter into the writing process, the level of fluency in Cluster 4 appears to drop notably, but goes back on and remains on that level throughout the remaining time of writing.

The DIRatio provides a measure of the extent of a writer's overall editing behavior. Deletion, in the current context, captures most kinds of editing that a writer conducts. Throughout the writing process, generally speaking, Cluster 1 shows more editing behavior than Cluster 2; Cluster 2 shows similar extent of editing to Cluster 3, although not with the same sequential pattern; and Cluster 4 shows the least extent of editing behavior of all clusters (Fig. 3). This result is evident based on both the intercept estimates (−1.918 for cluster 1, −2.084 for cluster 2, −2.226 for cluster 3, and −2.547 for Cluster 4 in Table 6) and the trend lines in Fig. 3. For both Clusters 1 and 2, there is the greatest number of deletion actions relative to insertion actions at the beginning of the writing process, with this relative deletion decreasing steadily as writing proceeds. Simple linear regressions detect a statistically significant linear trend for both Cluster 1 and Cluster 2 (Table 6). There is also a significant linear trend detected for Cluster 3, for which the slope estimate is close to that of Cluster 1, suggesting that the relative number of deletions to insertions is greater at the beginning of the writing and it becomes lower as time goes along. However, it is observed (second panel in Fig. 3) that Cluster 3 writers appear to conduct somewhat more deletion relative to insertion at the very end of the writing process. In contrast, the trend for Cluster 4 (the lowest scoring, struggling group) is again quite distinctive from that for the other three clusters. First, as mentioned earlier, Cluster 4 writers show overall the least extent of editing behaviors; and second, their editing trend across the writing process is not linear, but visually quadratic. Specifically, they appear to conduct more editing in the middle of the writing process, and less editing at the beginning and end of the writing process.

The number of keystrokes per second is an indication of a writer's general writing speed. Higher speed indicates greater fluency. It is not surprising that Cluster 1 showed notably the greatest speed and fluency, compared to Clusters 2, 3, and 4. Cluster 1 further reveals a linear increase in general speed as writing proceeds. The estimated slope parameter (0.005) is significant at $p \leq 0.5$ level. It is interesting to find that even though Cluster 2 has significantly higher average essay scores than Clusters 3 and 4, the general writing speed for Cluster 2 is slower than those two clusters. The estimated intercept of simple linear regression is 1.326 for Cluster 2, which is more than 0.2 points lower than Cluster 3 (1.585, Table 6) and more than 0.3 points lower than Cluster 4 (1.682). This phenomenon is also apparent in the bottom panel in Fig. 3 where the trend line is visibly lower for Cluster 2 than Clusters 3 and 4. When combined with the previously reported results, we speculate that Cluster 2 writers, compared to Clusters 3 and 4 writers, are more persistent writers, even though they are not necessarily faster or more adept at keyboarding. Additionally, all but Cluster 1 show a quadratic pattern on this general writing speed measure: higher speed at the beginning and end of writing, and lower speed in the middle, although the pattern is more visibly apparent for the two lower scoring groups – Clusters 3 and 4. This result is consistent with findings reported in Zhang, Hao, Li, & Deane (2016) and Zhu et al. (2019).

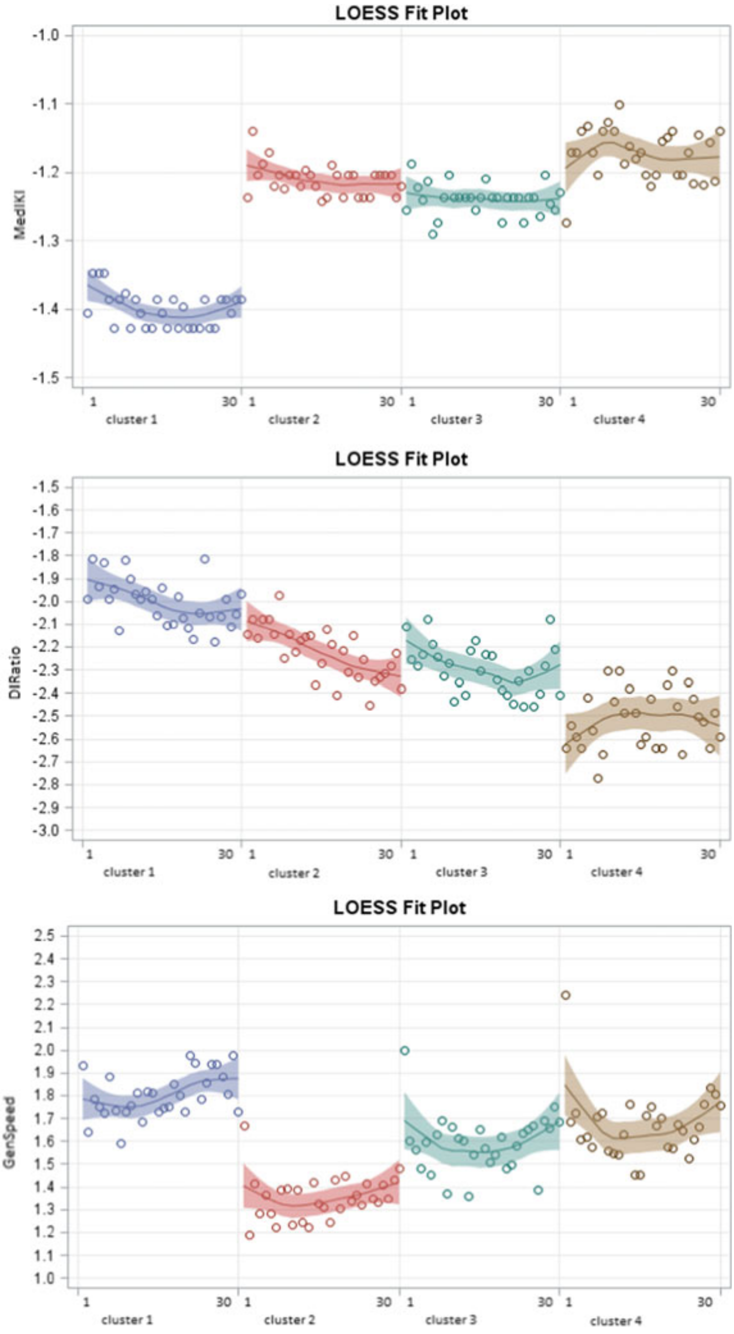


Fig. 3 Sequential patterns on writing characteristics across clusters. *Note* From top to bottom panels: MedIKI, DIRatio, and GenSpeed. Values 1–30 on the X-axis represent the 30 subsessions in each cluster

Table 7 Comparison of clusters based on editing summary features

Summary editing feature	Cluster 1 (n = 103)	Cluster 2 (n = 216)	Cluster 3 (n = 204)	Cluster 4 (n = 217)
AmountofDiscardedText	0.36^{a}	0.29^{b}	0.23^{c}	0.15^{d}
RateTypoCorrection	4.08^{a}	3.65^{b}	3.33^{c}	2.73^{d}
MedianJumpLength	110^{a}	67^{b}	81^{ab}	44^{c}
ProportionPreJumpPauseTime	0.09^{a}	0.08^{ab}	0.07^{ab}	0.06^{b}
MedMaxWordEditPauseTime	0.69^{a}	0.99^{ab}	0.96^{b}	1.09^{c}
ProportionEditedChunks	0.001^{ab}	0.001^{a}	0.001^{ab}	0.001^{b}
ProportionEditedWords	0.018^{a}	0.015^{b}	0.014^{b}	0.012^{c}
ProportionMinorEdits	0.019^{a}	0.017^{b}	0.017^{b}	0.016^{b}
ProportionMultiWordDelete	0.007^{a}	0.006^{ab}	0.006^{b}	0.005^{c}

Note A different superscription letter in each row suggests a significant mean difference at $p \leq 0.05$ within a column

Comparing Editing Behaviors. Finally, we compared the four clusters based on selected features that intended to characterize different kinds of editing behavior (Table 7). MANOVA results indicate a significant overall model effect: Pillai's Trace value is 0.64, with F(27.2049) = 20.54, $p < 0.0001$.

Generally speaking, Cluster 1 showed the greatest extent of various editing behaviors during the writing process. Specifically, compared to the other clusters, Cluster 1 as the most proficient group discarded more of their text relative to what was inserted during writing (which is consistent with the middle panel in Fig. 3); were more likely to correct typos immediately during their writing process; conducted text changes at places that were further from the previous cursor position; edited more words relative to how much they wrote while having shorter in-word pause time during the editing; and showed more small edits relative to how much they wrote. Cluster 4, as the lowest scoring group, showed the opposite pattern on all summary editing features compared to Cluster 1. Clusters 2 and 3 fell somewhere in the middle.

4 Discussion

There is a growing literature on the use of process data in digitally-delivered assessments (Ercikan and Pellegrino, 2017). This study represents our attempt to group students into different writing proficiency clusters and analyze students' editing behaviors in the identified clusters during their writing process. The results of this study can help inform the kinds of feedback that can be given to teachers and students. This study analyzed the sequential patterns over the course of writing on three process characteristics for the identified writer groups; namely, the median pause interval between keystrokes, the ratio of deletion over insertion, and the number of keystrokes produced per second. As a follow-up study to Zhu et al. (2019), each keystroke log sequence was evenly divided into 30 segments using total writing time, representing different stages of the writing process.

The results showed that, using four basic performance indicators, four meaningful clusters of writers could be identified. The four clusters differed significantly on their essay quality and writing processes. They differed on the mean essay score, mean total time spent on task, and mean total number of words in the final response. Specifically, the contrast between Clusters 2 and 3 is worth noting and discussing. These two clusters were statistically different on the above-mentioned three dimensions (i.e., score, time, and length) but not on their keyboarding skill as measured by the in-word typing speed. Results in RQ2 further revealed that Cluster 2 writers did not appear to be more fluent than cluster 3 writers, but showed greater writing effort and more engagement in the task than Cluster 3 writers. Cluster 2 writers spent more time on writing and edited more during the writing process, even though their typing skill was not statistically better and they appeared to be slower in general writing speed than Cluster 3 writers. One hypothesis is that motivation and general effort played a role in the detected differences between Clusters 2 and 3. A valuable follow-up study would be to evaluate this hypothesis. The four identified clusters also showed distinct sequential patterns over the course of writing on three process characteristics and, as well, differed on their editing behaviors during text production. The results of this study are largely consistent with previous findings that stronger writers tend to manage their writing more efficiently, produce text more fluently, and engage in more editing and revision. By contrast, weaker writers tend to produce text less efficiently and pause more frequently in locations that suggest difficulties in typing, spelling, word-finding and other transcription processes (Alves, Castro, de Sousa, & Stromqvist, 2007; Stevenson, Schoonen, & de Glopper, 2006). A final note is that the results of this study are limited to the writer population, which was comprised of U.S middle school students. As for a future study, it will be of value to examine whether the identified writer clusters hold across different writing genres and if the cluster assignment interacts with writers' demographic and social backgrounds.

References

Alves, R. A., Castro, S. L., de Sousa, L., & Stromqvist, S. (2007). Influence of typing skill on pauseexecution cycles in written composition. In M. Torrance, L. van Waes, & D. Galbraith (Eds.), *Writing and Cognition: Research and Applications* (pp. 55–65). Amsterdam: Elsevier.

Bennett, R. E., Deane, P., & van Rijn, P. W. (2016). From cognitive domain theory to assessment practice. *Educational Psychologist*, *51*, 82–107.

Cleveland, William S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association.*, *74*, 829–836. https://pdfs.semanticscholar.org/414e/5d1f5a75e2327d99b5bbb93f2e4e241c5acc.pdf.

Cooley, W. W., & Lohnes, P. R. (1971). *Multivariate data analysis*. John Wiley and Sons.

Deane, P., Feng, G., Zhang, M., Hao, J., Bergner, Y., Flor, M., Wagner, M., Lederer. N.: Generating scores and feedback for writing assessment and instruction using electronic process logs. US Patent and Trademark Office. Application No. 14/937,164 (2016).

Ercikan, K., & Pellegrino, J. W. (2017). *Validation of score meaning for the next generation of assessments: The use of response processes*. Taylor & Francis.

Guo, H., Deane, P., van Rijn, P., Zhang, M., & Bennett, R. (2018). Exploring the heavy-tailed key-stroke data in writing processes. *Journal of Educational Measurement*, *194–216*,

Murtagh, F., Legendre, P.: Ward's hierarchical clustering method: Clustering criterion and agglomerative algorithm. Accessed in October 2018: http://arxiv.org/abs/1111.6285.pdf (2011)

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithm implement Ward's criterion. *Journal of Classification*, *31*, 274–295.

Rencher, A. C. (1992). Interpretation of canonical discriminant functions, canonical variates, and principal components. *The American Statistician*, *46*, 217–225.

Sinharay, S., Zhang, M., Deane, P.: Application of data mining for predicting essay scores from writing process and product features. Applied Measurement in Education (2019). https://www.tandfonline.com/doi/full/10.1080/08957347.2019.1577245.

Stevenson, M., Schoonen, R., & de Glopper, K. (2006). Revising in two languages: A multi-dimensional comparison of online writing revisions in L1 and FL. *Journal of Second Language Writing*, *201–233*,

van Rijn, P., Yan-Koo, Y.: Statistical results from the 2013 CBAL English Language Arts multistate study: Parallel forms for argumentative writing. RM-16-15. Princeton, NJ: Educational Testing Service (2016).

van Rijn, P., Chen, J., Yan-Koo, Y.: Statistical results from the 2013 CBAL^{TM} English Language Arts multistate study: Parallel forms for policy recommendation writing. RR-16-01. Princeton, NJ: Educational Testing Service (2016).

Ward, J. H, Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*, 236–244.

Zhang, M., Bennett, R., Deane, P., & van Rijn, P. (2019). Are there gender differences in how students write their essays? An analysis of writing processes. *Educational Measurement: Issues and Practice, Online First*.

Zhang M., Deane, P.: Process features in writing: Internal structure and incremental value over product features. RR-15-27. Princeton, NJ: Educational Testing Service (2015).

Zhang, M., Feng, G., Deane, P., H, Guo.: Investigating an approach to evaluating keyboarding fluency. To be submitted for publication (2018).

Zhang, M., Hao, J., Li, C., & Deane, P. (2016). Classification of writing patterns using keystroke logs. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative Psychology Research*. New York: Springer.

Zhang, M., Hao, J., Li, C., Deane, P.: Defining personalized writing burst measures of translation using keystroke logs. In proceedings of the 2018 Educational Data Mining Conference, 549 - 552 (2018).

Zhu, M., Zhang, M., Deane, P.: Analysis of keystroke sequences in writing logs. RR-xx-xx. Princeton, NJ: Educational Testing Service (2019). https://onlinelibrary.wiley.com/doi/10.1002/ets2.12247.

Modeling Examinee Heterogeneity in Discrete Option Multiple Choice Items

Nana Kim, Daniel M. Bolt, James Wollack, Yiqin Pan, Carol Eckerly and John Sowles

Abstract A new format for computer-based administration of multiple- choice items, the discrete option multiple choice (DOMC) item, is receiving growing attention due to potential advantages related both to item security and control of test-wiseness. A unique feature of the DOMC format is the potential for an examinee to respond incorrectly to an item for different reasons—either failure to select a correct response, or selection of a distractor response. This feature motivates consideration of a new item response model that introduces an individual differences trait related to general proclivity to select response options. Using empirical data from an actual DOMC test, we validate the model by demonstrating the statistical presence of such a trait and discuss its implications for test equity in DOMC tests and the potential value for added item administration constraints.

Keywords Item response model · Discrete option multiple-choice items · Computer-based testing

1 Introduction

The multiple-choice (MC) item format is commonly used in large-scale educational tests due to its simple and objective scoring. However, weaknesses of the MC format such as a vulnerability to testwiseness and test security compromise threaten test validity by increasing construct-irrelevant variance. A new item format called discrete option multiple choice (DOMC) was recently suggested as a potential alternative (Foster & Miller, 2009).

N. Kim (✉) · D. M. Bolt · J. Wollack · Y. Pan
University of Wisconsin-Madison, Madison, WI 53706, USA
e-mail: nkim84@wisc.edu

C. Eckerly
Educational Testing Service, Princeton, NJ, USA

J. Sowles
Ericsson Inc., Santa Clara, CA, USA

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_33

The DOMC item, which is typically administered by computer, presents an item stem followed by a sequential and random presentation of response options rather than presenting all options at once as in MC items. For example, an item asking “Which determines the loudness of sound?” can have response options of (A) frequency (B) period and (C) amplitude, where (C) is the answer. In DOMC format, examinees are asked to choose “yes” or“no” to the question “Does this determine the loudness of sound?” with each option randomly presented. An item may also have more than one keyed option. The DOMC item is scored as correct if the examinee correctly endorses all keyed option(s), but is scored as incorrect if the examinee either fails to endorse a keyed option or endorses a distractor option. Because the administration of response options ends as soon as either an incorrect response is made or the last of the keyed response options is presented (except for the cases where an additional unscored response option is presented, as DOMC randomly assigns it with a fixed probability for purposes of masking the correctness of the final response), not all response options are presented to all examinees. Therefore, DOMC protects item security, and it also reduces the effect of testwiseness on test performance because it presents response options to examinees one at a time (Foster & Miller, 2009).

The purpose of this paper is to propose a new item response model that potentially provides a useful way to understand psychometric differences between DOMC and MC items. We speculate that when presented a DOMC item, one challenge faced by examinees will be a general uncertainty as to the degree of “correctness” for an individually presented response option, and whether it is sufficient to warrant endorsing it as a keyed response. How examinees deal with that uncertainty may vary; some may adopt a lower threshold in what is viewed as correct, others a higher threshold. The possibility of such an individual difference reflects a fundamental way in which DOMC items may perform differently than MC items where all examinees see all response options. Such an individual difference also ultimately affects the correctness of items and has the potential to contribute to inequity because the options are presented in random order in the DOMC format. A complete randomization of key location across items may lead to different average key location across examinees. Examinees who have a high tendency to pick options may be disadvantaged if their average key location is relatively late while those with a low tendency to pick would be disadvantaged if the average key location is earlier.

In this respect, we propose an item response model that introduces an individual differences trait related to a general proclivity to select response options. Using empirical data from an actual DOMC test, we validate the model by demonstrating the statistical presence of such a trait and discuss its implications for test equity in DOMC tests in relation to key location. Finally, we consider use of the model as a possible way to reduce the effect of this nuisance trait and key location on test scores.

1.1 An Item Response Model for DOMC Items

In our proposed model, we consider an examinee's responses to each presented response option within and across items by treating each response option as though they are separate "items". For a response option (j) presented to examinee (i), we model the probability of endorsing the option as:

$$P(U_{ij} = 1|\theta_i, \eta_i) = \frac{\exp(a_j\theta_i + b_j + \eta_i)}{1 + \exp(a_j\theta_i + b_j + \eta_i)} \tag{1}$$

where $U_{ij} = 1$ implies that examinee i selects option j if presented, θ_i represents an examinee's proficiency on the skill of interest, and η_i represents an examinee's tendency to select an option (whether it is keyed or distractor). Positive and negative values of η_i reflect tendencies to over-select and under-select presented options, respectively, irrespective of the correctness of those options. Table 1 illustrates an example of actual response patterns for examinees with a similar level of $\hat{\theta}_i$ but different levels of $\hat{\eta}_i$. The responses to options within each of the first 9 items presented to each examinee are shown (items 1 through 9 for one examinee are not identical to the items for the other because items are randomly administered in DOMC format). The responses of 0 and 1 respectively represent the rejection and selection of the option, and dots represent the options that are not presented to the examinees (due to the fact that options are no longer presented once the examinee gets the item correct or incorrect). The responses in the table show that the examinee with high $\hat{\eta}_i$ tends to endorse more often than the examinee with low $\hat{\eta}_i$ while the examinee with low $\hat{\eta}_i$ tends to reject options more frequently than the examinee with high $\hat{\eta}_i$. The proportions of selection (P(1)) and rejection (P(0)) also indicate that the examinee of high $\hat{\eta}_i$ endorsed most of the response options presented (about 76%) whereas the examinee of low $\hat{\eta}_i$ endorsed only about 36% and rejected 64% of the response options presented. Interestingly, this difference in response behavior is apparent despite the fact that the overall performance on the test is approximately the same ($\hat{\theta}_i$ = $-.46$, $-.48$) for the two respondents.

The parameters a_j and b_j in equation (1) respectively represent discrimination and difficulty parameters for option j. The a_j estimates are positive for keyed and negative for distractor options while the b_j estimates are positive for more frequently selected options and negative for less frequently selected options. An example of these parameter estimates for options within items are presented in Table 2. Items 1, 7, and 21 each has one, two and three keyed options, and items 1 and 7 have four

Table 1 An example of actual responses for different η_i levels

$\hat{\theta}_i$	$\hat{\eta}_i$	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	P(1)	P(0)
−0.46	0.91	01 ··	1 · · ·	111 ·	1 · · ·	10 ··	11 ··	1 · · ·	01 · · ·	1 · · ·	0.76	0.24
−0.48	−0.60	1· · ·	000 ·	0 · · ·	000 ·	0 · · ·	0000	01 ··	110 ··	01 ··	0.36	0.64

Table 2 An example of item estimates for one, two, and three keyed items

Item	option 1		option 2		option 3		option 4		option 5	
number	$\hat{a}_j$	$\hat{b}_j$	$\hat{a}_j$	$\hat{b}_j$	$\hat{a}_j$	$\hat{b}_j$	$\hat{a}_j$	$\hat{b}_j$	$\hat{a}_j$	$\hat{b}_j$
1	1.49	0.98	−1.23	−1.92	−1.14	−1.23	−1.93	−1.90	·	·
7	0.46	0.53	0.14	0.39	−1.21	0.11	−0.90	0.18	·	·
21	0.88	2.50	0.99	1.99	0.38	0.16	−1.02	−1.23	−1.88	−1.21

possible options while item 21 has five possible options. Noting that option 1 is a keyed option for item 1, options 1 and 2 are keyed for item 7, and options 1, 2, and 3 are keyed for item 21, we can observe that a_j estimates for keyed options are positive and those for distractor options are negative. Such a property of a_j having both positive and negative values separates θ_i from η_i which has constant loadings of 1 across all options and items.

2 Methods

The data we use are operational data from an information technology certification test delivered in the DOMC format. The dataset has 648 examinees and a pool of 83 items (two test forms with 59 items each) where 54 items are single-keyed items, 24 items are two-keyed items, and 5 items are three-keyed items. Each examinee is administered one of the 59-item forms. Items with one keyed option have three incorrect response options, whereas all items with multiple keyed options have two incorrect response options.

We fit the model using a Bayesian (Markov chain Monte Carlo) estimation algorithm using WinBUGS1.4 (Lunn, Thomas, Best, & Spiegelhalter, 2000). For priors of the option parameters, we assume $a \sim Normal(0, 1)$ and $b \sim Normal(0, 1)$. For the examinee parameters, we assume $\theta \sim Normal(0, 1)$ and $\eta \sim Normal(0, var)$ where $1/var \sim Gamma(1, 1)$.

We seek to validate the model in two ways. First, we compare model fit with a model that excludes η_i using the Deviance Information Criterion (DIC) (Spiegelhalter, Best, Carlin, & van der Linde, 2002). Second, we examine the predictive effects of η_i on the resulting test score. Specifically, using η_i estimates from the WinBUGS analysis, we examine the effects of $\hat{\eta}_i$ and key location on the test score using regression analysis. We define loc_i to be the average scheduled location of the last key options across items for the examinee, a value determined from the randomly assigned schedule of response options for the examinee. Then we fit

$$X_i = \beta_0 + \beta_1\hat{\theta}_i + \beta_2\hat{\eta}_i + \beta_3 loc_i + \beta_4\hat{\eta}_i loc_i + e_i, e_i \sim N(0, \sigma^2) \qquad (2)$$

where X_i is a total sum score for examinee i, loc_i is the average last key location for all items presented to examinee i, and β_4 represents the interaction effect between $\hat{\eta}_i$ and loc_i. Each of the variables ($\hat{\theta}_i$, $\hat{\eta}_i$, and loc_i) were mean centered.

To evaluate implications for test inequity, we examine the distributions of true scores for a hypothetical examinee at a specific level of θ and η from 1000 hypothetical administrations of the DOMC test. We compare the resulting true score distributions under two hypothetical test administration conditions, one based on a constrained and one based on a complete randomization of key location. For a constrained randomization condition, we fix the distribution of key locations across examinees and randomize under that constraint. Specifically, we constrain 9, 10, 10, and 9 single-keyed items to have key locations of 1, 2, 3, and 4, respectively; 5, 6, and 6 two-keyed items to each have last key locations of 2, 3, and 4; and 1, 2, and 1 three-keyed items to have the last key locations of 3, 4, and 5, respectively. In current practice, a complete randomization approach is used. We seek to demonstrate how a constrained randomization approach may help protect against test inequity.

3 Results

3.1 Validation of the η_i Trait

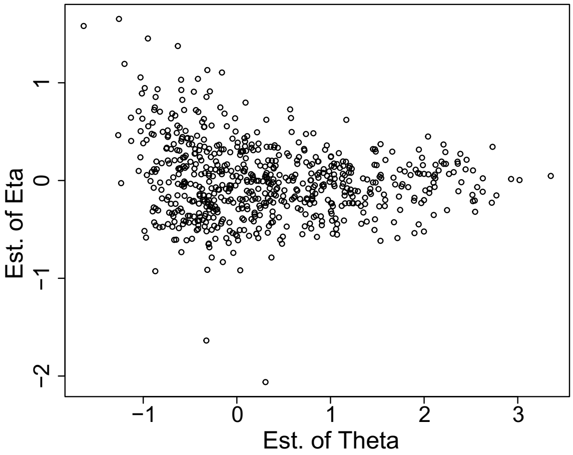
Using the Deviance Information Criterion (DIC), we find the model in equation (1) provides a better comparative fit to the empirical data than a model that excludes η_i (Table 3).

We also examined a scatterplot of the θ_i and η_i estimates. One of the features of the $\hat{\eta}_i$ trait is that its variability is greatest among low $\hat{\theta}_i$ examinees (see Fig. 1). This result illustrates how the tendency to over- or under-select options could be seen as a factor contributing to low ability estimates on DOMC tests. In this respect, the η_i estimate also provides diagnostic information, as poor performance of the test may often be due to tendencies to either over-select or under-select among response options. Also, the correlation between $\hat{\eta}_i$ and $\hat{\theta}_i$ turns out to be -0.146 which is close to zero, indicating that a tendency to over- or under-select options does not necessarily correlate with an examinee's proficiency.

Table 3 A comparison of model fit

Models	Dbar	Dhat	pD	DIC
without η_i	94424.6	93166.9	1257.8	95682.4
with η_i	92271.5	90496.4	1802.1	94073.5

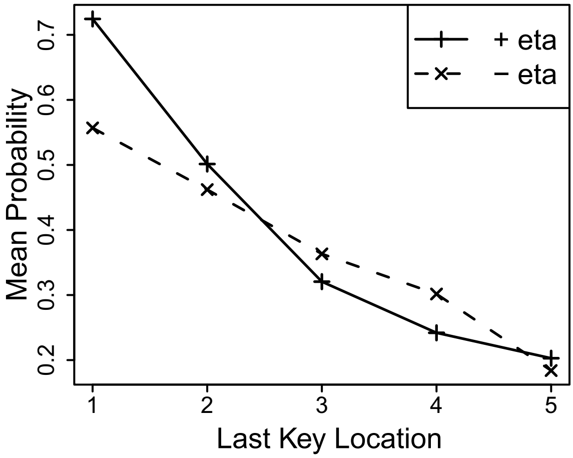
Fig. 1 A plot of η_i estimates by θ_i estimates



3.2 Effects of $\hat{\eta}_i$ and Key Location on Test Scores

Test scores for examinees with the same θ_i but different η_i can differ due to random differences in the average scheduled key location. As noted earlier, examinees of extreme positive η_i will be disadvantaged by later average key locations while those with extreme negative η_i will be disadvantaged by earlier average key locations. Figure 2 shows how item difficulty changes differently in relation to key location for examinees of positive $\hat{\eta}_i$ versus negative $\hat{\eta}_i$. The graph reports the empirically estimated item difficulty (item-level p-values) comparing examinees with either positive or negative $\hat{\eta}_i$. The lines for examinees of positive and negative $\hat{\eta}_i$ are reversed as the key location changes. Specifically, we can observe that early key locations

Fig. 2 Average item difficulty by key location for positive and negative $\hat{\eta}_i$



make items more difficult for examinees of negative $\hat{\eta}_i$ who tend to under-select response options, whereas late key locations make items more difficult for examinees of positive $\hat{\eta}_i$.

Table 4 shows the result of our regression of test score on $\hat{\eta}_i$, the average last key location, their interaction, and the examinee proficiency estimates. As expected, loc_i has a negative effect, meaning a later average key location makes the test more difficult. In addition, the negative interaction implies that a more positive $\hat{\eta}$ strengthens this effect, making a later key location particularly disadvantageous for examinees of high $\hat{\eta}$. This also can be interpreted as implying that an earlier key location can be advantageous for examinees of high $\hat{\eta}$ while disadvantageous for examinees of low $\hat{\eta}$.

The graphs in Fig. 3 illustrate the interaction effect between $\hat{\eta}_i$ and key location for low, average, and high ability groups. They suggest that the effects of $\hat{\eta}_i$ can yield different levels of bias depending on the average key location. The bias seems to be larger in positive $\hat{\eta}_i$ levels, and up to ± 10 score points of bias can be created. Specifically, for examinees with a high level of $\hat{\eta}_i$, those who are presented with early and late key locations can respectively obtain scores 10 points higher and lower than those who are presented items at the mean key location.

Table 4 Estimated effects of variables on test scores

	Coef.	S.E.	*t*	*p*-value
Const.	19.69	0.09	208.66	0.000
Theta($\hat{\theta}_i$)	10.26	0.10	105.60	0.000
Eta($\hat{\eta}_i$)	2.22	0.23	9.56	0.000
Key loc.	−7.51	0.70	−10.76	0.000
Eta x Key loc.	−7.89	1.80	−4.39	0.000

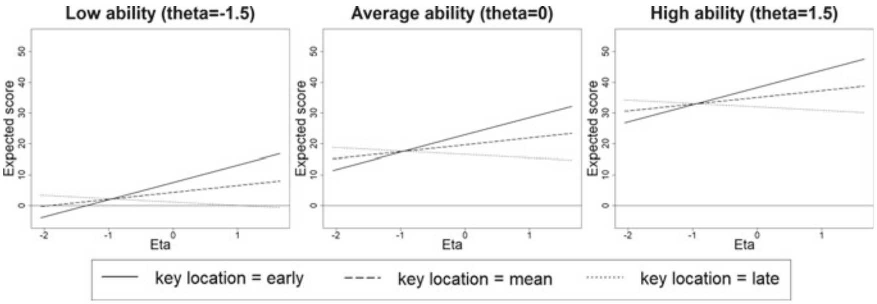


Fig. 3 Interaction effects of $\hat{\eta}_i$ and key location

3.3 Constrained Administration of DOMC Items

The distributions of true scores under complete and constrained randomization of key location were compared. The box plots in Fig. 4 illustrate the distributions of true scores for examinees of $\theta_i = -0.5$, 0.5, and $\theta_i = 1.5$ at different levels of η_i for two test forms (f1 and f2). The distributions of true scores under the complete randomization condition (which is currently applied for DOMC items) has a large variability, and the variance is substantially reduced under the constrained randomization condition. This indicates that the effect of key location (including the interaction effect between η_i and key location) is reduced when the distribution of key locations is constrained to be equal across examinees. Such results suggest that a constrained randomization of key location may significantly reduce test inequity effects of η_i that currently exist.

4 Conclusions

Our results confirm a distinguishable tendency for examinees to over- and under-select response options in DOMC items as captured by the η_i parameter. An additional goal in proposing a model attending to these effects is to study how η_i may interact with the random administration of response options to increase variability in the effects of η_i across administrations. Because the random administration of response options does not guarantee equal average key location across examinees, examinees' different tendencies to endorse or not endorse response options when uncertain about the options can create bias in test scores, which leads to a test inequity issue. Our study shows that such an interaction can be controlled through a constrained randomization procedure that ensures a consistent distribution of last key location across administrations, a constraint that should seemingly be easy to implement and at little cost to test integrity. By controlling the distribution of key locations, we can fix the average last key location across examinees and reduce the effects of η_i due to different average key locations. Finally, the model may also serve a diagnostic purpose, particularly for lower-ability examinees, in that estimates of η_i can help identify respondents that are miscalibrated in terms of their response behavior in the presence of uncertainty.

There are a number of potential directions for future work with this model and/or DOMC items. For example, one question is whether examinee estimates of η_i remain stable across administrations. It might be speculated that the variability in η_i should diminish with increased examinee experience with DOMC items, as the format is still relatively new to most respondents. Along these lines, it may even be useful to study whether η_i changes within a single test administration. Given the length of the current test, it is conceivable that examinees learn to adjust their response behavior in productive ways before they reach the end of the test. Another area of potential interest concerns item effects. Our analyses assume a consistency in the effects of

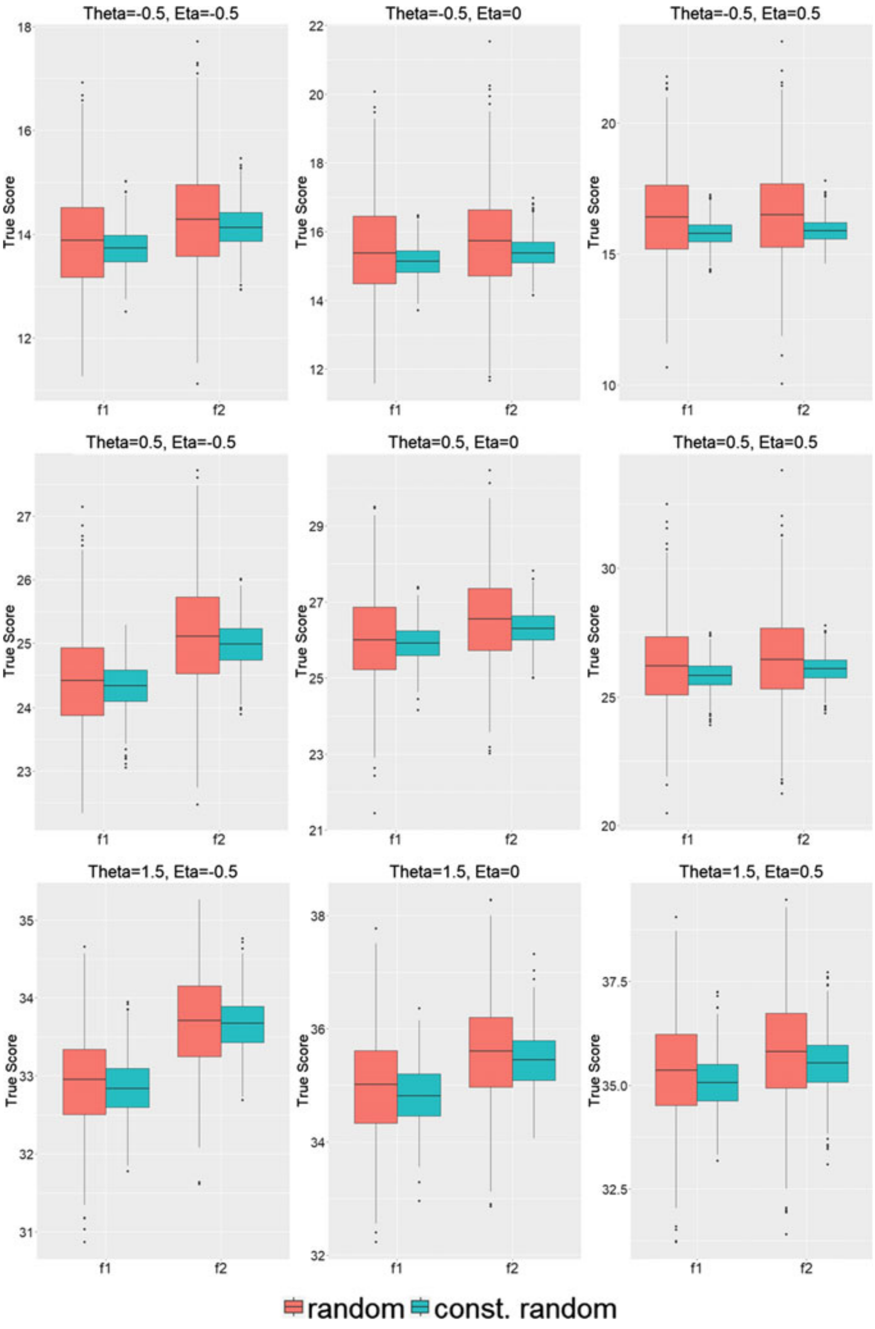


Fig. 4 Box plots of true scores for complete and constrained randomization of key location

η_i across items. It might be explored whether this is in fact the case, or whether the behavior is more common with certain items or response options than others. Moreover, we can also examine whether the effect of η_i changes across examinees' ability levels. Though we did not consider the η_i effect to vary across ability levels in our regression analysis, it is probable that examinees of low ability are more largely affected by η_i (as seen from a larger variability of η_i for low θ_i levels in Fig. 1) because they encounter an uncertainty more often. Another area of future research should attend to methods for evaluating the absolute fit of the model. One complexity in applying the types of posterior predictive checking (PPC) methods commonly used with Bayesian estimation methods relates to the restrictions imposed on the item response data structure due to the use of the DOMC format. Specifically, within a common item, correct or incorrect responses to certain options will preclude the ability to observe responses to other options within the same item (which will not be administered as a result of the DOMC format). This problem seemingly applies to all psychometric models that might be applied under the DOMC administration format and may require creative solutions. Finally, the results of our model might be informed by actual interviews with examinees administered the DOMC items. It is unclear whether the variability we see is related to different beliefs about the relative likelihoods of keyed/non-keyed response options on the test or is actually driven by behavior that is inconsistent with probability theory.

References

Foster, D., & Miller H. L. (2009). A new format for multiple-choice testing: Discrete-Option Multiple-Choice. Results from early studies A new format for multiple-choice testing: Discrete-option multiple-choice. results from early studies. *Psychology Science Quarterly*, *51*(4), 355–369.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. J. (2000). WinBUGS - A Bayesian modeling framework: Concepts, structure and extensibility Winbugs - a bayesian modeling framework: Concepts, structure and extensibility. *Statistics and Computing*, *10*(4), 325–337. https://doi.org/10.1023/A:1008929526011.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fitBayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353.

Simulation Study of Scoring Methods for Various Multiple-Multiple-Choice Items

Sayaka Arai and Hisao Miyano

Abstract Multiple-choice (MC) format is the most widely used format in objective testing. The "select all the choices that are true" items, also called multiple-multiple-choice (MMC) items, is a variation of the MC format, which gives no instructions about how many correct choices may be selected. Although many studies have been developed and various scoring methods for MMC items have been compared, the results have often been inconsistent. Arai and Miyano (Bull Data Anal Japan Classif Soc 6:101–112, 2017) proposed new scoring methods and compared their scoring features by conducting numerical simulations of a few MMC item patterns. In this study, we conducted numerical simulations of all other plausible MMC item patterns to examine the relationships between examinees' abilities (true scores) and scores given by scoring methods. We illustrated the effects of the total number of choices and correct choices for each scoring.

Keywords Multiple-multiple-choice items · Scoring method

1 Introduction

Multiple-choice format (MC) is the most widely used format in the field of objective testing. It provides several choices and usually gives only one correct choice of answers. In a typical MC format, examinees choose one answer. The scoring method is a binary one that gives 1 point for a correct answer and 0 points for a wrong answer; there is no middle ground between them.

However, there are also formats that have multiple correct choices. In such formats, partial points can be given in accordance with the choices examinees select. For example, if the number of correct choices is indicated, it is possible to give a partial point in an easy-to-understand manner such as making it proportional to the number

This work was supported by JSPS KAKENHI Grant Number JP18K03085.

S. Arai (✉) · H. Miyano
National Center for University Entrance Examinations, Tokyo, Japan
e-mail: sayarai@rd.dnc.ac.jp

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_34

of correct choices examinees have selected. On the other hand, if the number of correct choices is not indicated, the number of correct choices becomes a part of the question to be answered, so partial points cannot be given in such an easy-to-understand way.

1.1 Multiple-Multiple-Choice (MMC) Items and Partial Points

An MC format with no instructions on the number of correct choices is a variation of the MC format. It is called “multiple-multiple-choice format”,“multiple-mark format”, or “type X form.” It has also been categorized as a variation of the“multiple true-false form” (Haladyna, 2004). In this study, we call it multiple-multiple-choice (MMC) format.

A typical MMC format question is “select all answers that are true.” In the MMC format, examinees need to judge correctness of all of the presented options. Therefore, this format is said to be an effective way for measuring detailed knowledge about a specific field. An example of the MMC format items is shown below (Tsai & Suen, 1993).

MMC format example
As the sample size increases (select all that are true),
A. the sampling distribution of the mean is more like . . .
B. the t-distribution is more like a normal . . .
C. the number of degrees of freedom . . .
D. . . .

Although it is possible to give partial points in the MMC format, it is not simple to decide how many points should be given as partial points. Let’s consider the case of an item that has four choices with correct choices A and B. If the examinee’s answer is “A and B”, the full score will be given. If the examinee’s answer is “A only” or “A, B and C”, partial points should be given, but which answer should be given a higher score? Or, if the examinee’s answer is “A, B, C, and D (all choices)”, should partial points be given because the correct choices are included in the answer?

1.2 Scoring Method for MMC Format Items

Scoring methods for the MMC format have been studied for a long time (for example, Cronbach, 1941). Although many studies have developed and compared scoring methods for the items of this type, the results have often been inconsistent, except that giving partial points increases the reliability and validity of the test (Albanese & Sabers, 1988; Domnich et al., 2015; Tsai & Suen, 1993).

Several scoring methods have been proposed so far. In what follows, we describe five methods for which N is the total number of choices, n_c is the number of correct choices, and n_i is the number of incorrect choices ($N = n_c + n_i$). Here, at least one correct choice is assumed to be included, and the same is true for incorrect choices ($0 < n_c$, $0 < n_i$). Also, it is assumed that examinees select x_c choices from the correct choices and x_i choices from the incorrect choices ($0 \leq x_c \leq n_c, 0 \leq x_i \leq n_i$). Then it follows that the number of choices selected by examinees is $x_c + x_i$ and the number of correct responses (the number of matched pairs) m is $m = x_c + (n_i - x_i)$ ($0 \leq m \leq N$). Some of these notations are summarized in Table 1 for reference.

Multiple-response (MR) method Only if all the responses are correct ($m = N$), 1 point is given, otherwise it is 0 points. This method doesn't give any partial points.

$$s_{MR} = \begin{cases} 1 \ (m = N) \\ 0 \ (otherwise) \end{cases}$$

Count for n options correct (C1) method If all the responses are correct ($m = N$), 1 point is given, if there is only one wrong response ($m = N - 1$), 0.5 points are given, otherwise it is 0 points.

$$s_{C1} = \begin{cases} 1 & (m = N) \\ 0.5 & (m = N - 1) \\ 0 & (otherwise) \end{cases}$$

Multiple true-false (MTF) method Consider each choice as an item for a true-false item and give a partial point proportional to the number of correct responses. This method has an easy-to-understand scoring policy and was used in many previous studies.

$$s_{MTF} = \frac{m}{N}$$

Jaccard coefficient (Jac) method This method was recently proposed by Arai and Miyano (2017). It is based on the similarity between the response patterns and the key patterns. Negative matches are excluded from the numerator. Among the indices that do not include such a "negative match" in the similarity, the Jaccard coefficient is the simplest one (Sokal & Sneath, 1963). The range of the score is $0 \leq s_{Jac} \leq 1$.

$$s_{Jac} = \frac{x_c}{n_c + x_i}$$

Table 1 Notation

	Examinee		
	Select	Not select	Total
Correct choices	x_c	$n_c - x_c$	n_c
Incorrect choices	x_i	$n_i - x_i$	n_i
Total	$x_c + x_i$	$N - (x_c + x_i)$	N

Negative marking (NM) method In this method, a penalty is given when an examinee selects an incorrect choice. $\frac{1}{n_c}$ points are given when a correct choice is selected and $\frac{1}{n_i}$ points are deducted when an incorrect choice is selected. We call this method the "negative marking (NM) method" in this study. Giving scores proportional to the number of correctly selected choices is the same as that used in the Ripkey method (Ripkey, Case, & Swanson, 1996), but the process performed when incorrect choices are selected is different. The score will be a negative value when $x_c < x_i$, so we set $s_{NM} = 0$ when $s_{NM} < 0$.

$$s_{NM} = \begin{cases} \frac{x_c - x_i}{n_c} & (x_c \geq x_i) \\ 0 & (otherwise) \end{cases}$$

1.3 Purpose of Study

Although many studies have developed and compared scoring methods for the MMC items, the results have often been inconsistent. Arai and Miyano (2017) proposed new scoring methods and compared their scoring features using numerical simulations. However, they showed only a few examples.

In this study, we conducted numerical simulations of other MMC item patterns and examined the relationships between the examinees' abilities (true score θ) and scores given by scoring methods. We illustrated the effects of the total number of choices and correct choices on each scoring method.

2 Comparing Scoring Methods on the Basis of Desirable Properties

2.1 Desirable Properties as Partial Points

Partial points should reflect the abilities of examinees properly; that is, it would be desirable that partial points satisfy the following properties.

- Large number of steps to divide the scores finely;
- Easy-to-understand meaning of partial points, in particular, the meaning of 0 points;
- Low scores for guessing;
- No scores for selecting none or all choices.

2.2 Properties of Scores and Their Illustrations

We illustrated the effects of the total number of choices and correct choices on each scoring method (Figs. 1 and 2).

Figure 1 shows the number of partial point steps. The Jac method has more steps than other methods. The Jac method has many steps when n_c is large.

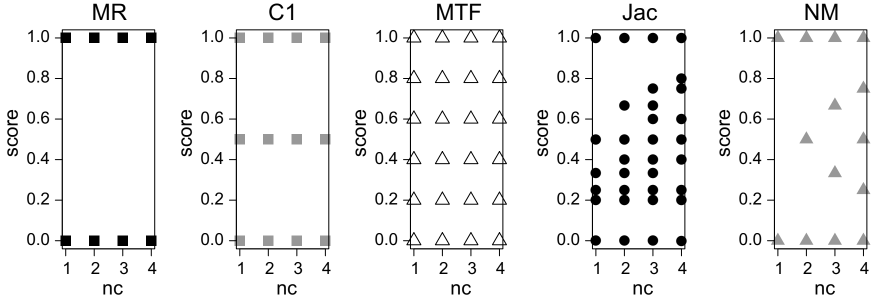


Fig. 1 Number of partial point steps ($N = 5$)

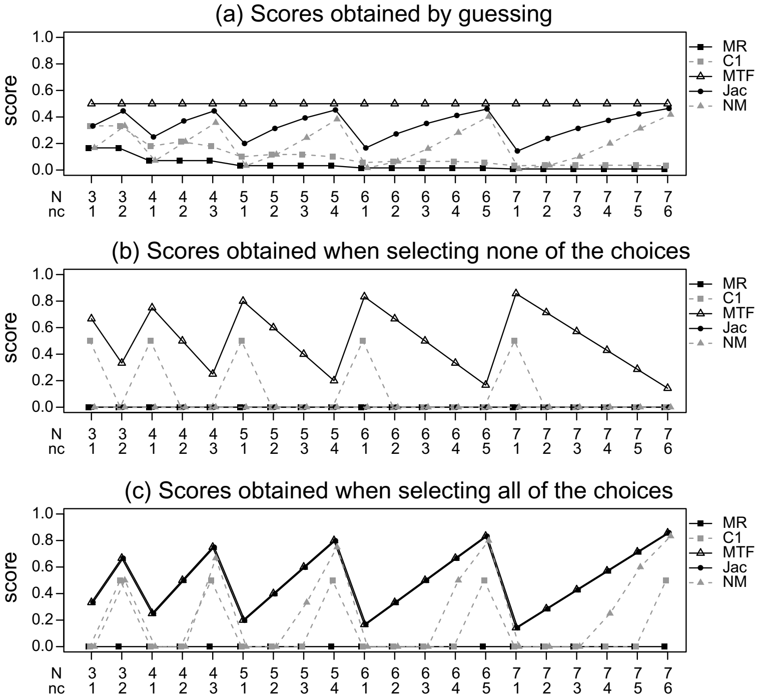


Fig. 2 Partial points in three extreme cases

Figure 2 shows the partial points given in extreme cases. The total number of choices (N) varied from 3 to 7. The number of correct choices (n_c) varied from 1 to $(N - 1)$. Figure 2a shows the partial points given when the examinees selected their choices by guessing. The scores shown here are the average values obtained when the examinees made their choices completely at random. In the MTF method, the expected values were 0.5 constantly. In the Jac and NM methods, the expected values increased as n_c increased. In the MR and C1 methods, the expected values were low in most cases when N was larger than 3. Figure 2b shows the partial points given when the examinees did not select any choices. In the MR, Jac, and NM methods, the scores were 0 constantly. In the C1 method, 0.5 score was given when $n_c = 1$. On the other hand, in the MTF method, a score was given if the examinee did not select any choices. Figure 2c shows the partial points given when the examinees selected all the choices. In this case, the scores were constantly 0 only for the MR method. In particular, when $n_c = N - 1$, 0.5 or more points were given for every method except the MR method. In the MTF and Jac methods, the same points were scored, and the scores were higher than in other methods.

We calculated the properties of each score from the definition formula for each scoring method and summarized them in Table 2.

Table 2 Score properties

Properties	s_{MR}	s_{C1}	s_{MTF}	s_{Jac}	s_{NM}
Range of scores	0 or 1	0, 0.5, or 1	[0, 1]	[0, 1]	[0, 1]
Meaning of 0 points	$m \neq N$, i.e., none of the choices selected are correct	$m < N - 1$, i.e., the number of incorrect responses is 2 or more	$m = 0$, i.e., all the responses are incorrect	$x_c = 0$, i.e., no choices selected from correct choices	$x_c = x_w$, i.e., the number of selected correct choices is the same as the number of selected incorrect choices
Meaning of 1 point	$m = N$, i.e., all the choices selected are correct and all the choices not selected are incorrect				
Number of steps	2	3	$N + 1$	$1 + n_c(n_w + 1)$ at most	$N + 1$
No choices	0 points	0.5 points ($n_c = 1$), 0 points (otherwise)	$\frac{n_w}{N}$ points	0 points	0 points
All choices	0 points	0.5 points ($n_c = N - 1$), 0 points (otherwise)	$\frac{n_c}{N}$ points	$\frac{n_c}{N}$ points	$1 - \frac{n_i}{n_c}$ points

3 Simulations

We conducted numerical simulations of MMC items with N choices and n_c correct choices ($3 \leq N \leq 7, 1 \leq n_c \leq N - 1$) to illustrate the effects of the total number of choices and the correct choices on each scoring method.

3.1 Simulation Methods

We assumed that each choice independently follows a one parameter logistic model (1 PLM) with its difficulty parameter being zero. We assumed that the examinees' θ (true scores) were $-2.7, -2.1, -1.5, \ldots, 2.1, 2.7$ (10 groups) and that there were 100 examinees in each group. Therefore, the total number of examinees was $100 \times 10 = 1000$. We generated 1000 examinees' response patterns using the R package lazy.irt (Mayekawa, 2018) and calculated scores using the five scoring methods.

For example, for $N = 5$ and $n_c = 2$, we regarded five choices as five 1 PLM items and generated five 0/1 item responses per examinee. We set the first two choices as correct answers, i.e., "11000" is the key response (answer). An examinee's response "10100" means $x_c = 1$ and $x_i = 1$, so the examinee's score is 0 in the MR method, 0 in the C1 method, 0.6 in the MTF method, 0.33 in the Jac method, and 0 in the NM method.

We calculated Spearman's rank correlation coefficient in each simulation as an index representing the relationship between the ability values and the scores.

3.2 Simulation Results

Figure 3 shows the simulation results when $N = 5$. The size of each filled circle reflects the frequency of scores in each examinees' group of ability θ. In the MR, C1, and MTF methods, the relationship of partial points and ability were irrelevant to the number of correct choices (n_c). But in the Jac and NM methods, the number of examinees obtaining partial points increased with n_c.

The results obtained with the Spearman's rank correlation coefficients are shown in Fig. 4: in each case, the correlation coefficients in the MTF method were the highest and those in the MR method were the lowest. Correlation coefficients in the NM method were as low as those in the MR method when $n_c = 1$.

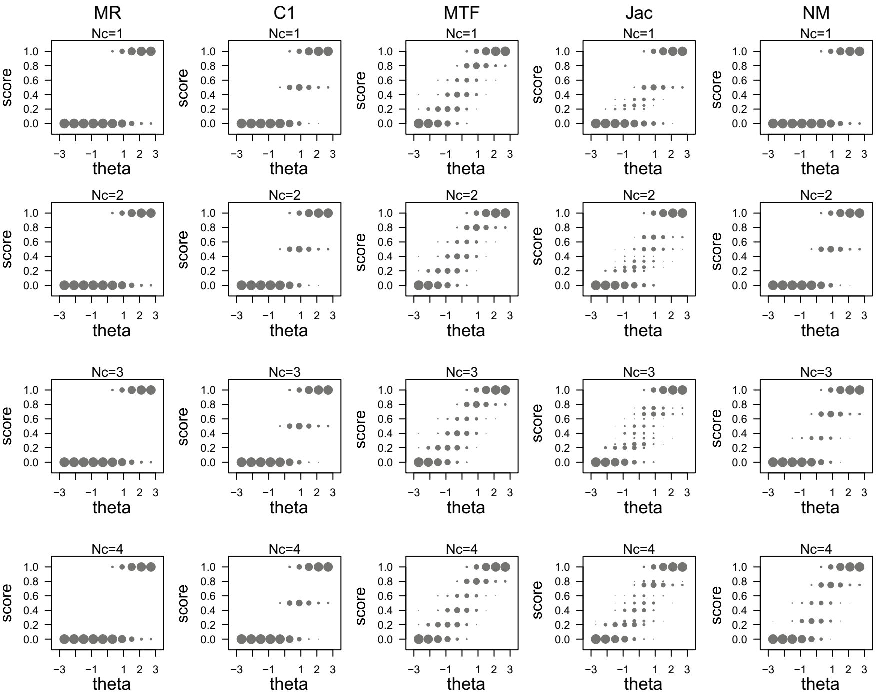


Fig. 3 Scatter plots of scores and θ ($N = 5$)

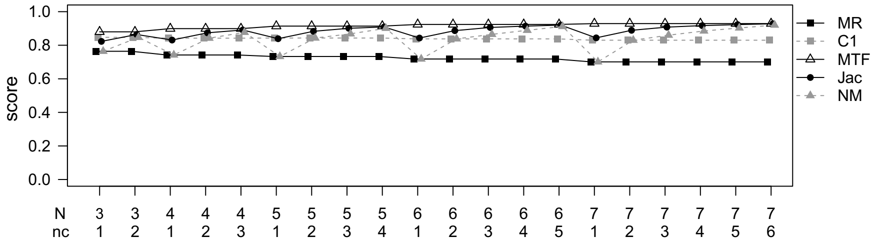


Fig. 4 Spearman's rank correlation coefficient

4 Discussion

We compared five scoring methods for MMC items from the viewpoint of desirability. Widely varying scores were obtained with the Jac method, which may reflect the fact that it gives a large number of partial point steps. The scores obtained by guessing for the MTF method were always 0.5. They were also high for the Jac and NM methods, with large n_c. The MTF method gave partial points when no choices were selected and all the scoring methods, except the MR method, gave partial points when all the choices were selected.

Scores obtained by guessing or by selecting none or all of the choices should be low. In the MTF method, scores obtained by guessing were 0.5 and those obtained by selecting no choices were also high. Although the simulation study results indicated that scores in the MTF method were relevant to θ, the MTF method is not a recommended method. Both the Jac method and the NM method were even better than the MTF method, although the scores obtained by selecting all the choices were rather high in both methods. On the other hand, the MR and C1 methods did not have any such embarrassing features. However, the simulation results showed that the MR method scores were less relevant to θ.

No single scoring method is optimal; rather, various factors need to be taken into consideration. These include the purpose of the examination and the policy on giving partial points, i.e., how many points are given to what kind of answers.

In this study, we compared five possible scoring methods for MMC items and clarified their characteristics. It is our hope that the results we have described in this paper will prove to be beneficial to other researchers doing work in this field.

References

Albanese, M. A., & Sabers, D. L. (1988). Multiple true-false items: A study of interitem correlations, scoring alternatives, and reliability estimation. *Journal of Educational Measurement*, *25*(2), 111–123.

Arai, S., & Miyano, H. (2017). Scoring method for “Select All the Choices That Are True” items. *Bulletin of Data Analysis of Japanese Classification Society, 6*(1), 101–112(in Japanese).

Cronbach, L. J. (1941). An experimental comparison of the multiple true-false and multiple multiple-choice tests. *Journal of Educational Psychology*, *32*(7), 533.

Domnich, A., Panatto, D., Arata, L., Bevilacqua, I., Apprato, L., Gasparini, R., et al. (2015). Impact of different scoring algorithms applied to multiple-mark survey items on outcome assessment: An in-field study on health-related knowledge. *Journal of Preventive Medicine and Hygiene*, *56*(4), E162.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mayekawa, S. (2018, November 15). *Lazy R packages*. Retrieved from http://mayekawa.in.coocan.jp/Rpackages.html.

Ripkey, D. R., Case, S. M., & Swanson, D. B. (1996). A “new” item format for assessing aspects of clinical competence. *Academic Medicine*, *71*(10), S34–6.

Sokal, R. R., & Sneath, P. H. A.(1963). *Principles of numerical taxonomy*. W. H. Freeman and Company.

Tsai, F. J., & Suen, H. K. (1993). A brief report on a comparison of six scoring methods for multiple true-false items. *Educational and Psychological Measurement*, *53*(2), 399–404.

Additive Trees for Fitting Three-Way (Multiple Source) Proximity Data

Hans-Friedrich Köhn and Justin L. Kern

Abstract Additive trees are graph-theoretic models that can be used for constructing network representations of pairwise proximity data observed on a set of N objects. Each object is represented as a terminal node in a connected graph; the length of the paths connecting the nodes reflects the inter-object proximities. Carroll, Clark, and DeSarbo (J Classif 1:25–74, 1984) developed the INDTREES algorithm for fitting additive trees to analyze individual differences of proximity data collected from multiple sources. INDTREES is a mathematical programming algorithm that uses a conjugate gradient strategy for minimizing a least-squares loss function augmented by a penalty term to account for violations of the constraints as imposed by the underlying tree model. This article presents an alternative method for fitting additive trees to three-way two-mode proximity data that does not rely on gradient-based optimization nor on penalty terms, but uses an iterative projection algorithm. A real-world data set consisting of 22 proximity matrices illustrated that the proposed method gave virtually identical results as the INDTREES method.

Keywords Additive trees · Three-way data · Individual differences · Iterative projection

1 Introduction

Additive trees are graph-theoretic models that can be used for constructing network representations of pairwise proximity data observed on a set of N objects (see, e.g., Sattath & Tversky, 1977). Each object is represented as a terminal node in a connected graph; the length of the paths connecting the nodes reflects the inter-object proximities. Least-squares methods have been proposed for fitting additive trees to $N \times N$ (two-way one-mode) proximity matrices such that the sum of the squared discrepancies between the observed proximities and the corresponding estimates of

H.-F. Köhn (✉) · J. L. Kern
Department of Educational Psychology, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA
e-mail: hkoehn@illinois.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_35

the path lengths or tree-distances is minimized. Carroll et al. (1984; see also, De Soete & Carroll, 1989) developed INDTREES, an algorithm for fitting additive trees to three-way two-mode proximities as they may be collected from multiple data sources within the context of cross-sectional and longitudinal studies. Individual differences are modeled explicitly by constructing separate, source-specific additive trees. INDTREES is a mathematical programming algorithm that uses a conjugate gradient strategy for minimizing a least-squares loss function augmented by a penalty term to account for violations of the constraints as imposed by the underlying tree model.

In this article, an alternative method for fitting additive tree structures to three-way two-mode proximity data is proposed that uses an iterative projection algorithm (Deutsch, 2001; Dykstra, 1983; Hubert & Arabie, 1995; Hubert, Arabie, & Meulman, 2006) and does not rely on gradient-based optimization nor on penalty terms for minimizing a constrained least-squares loss function. Comparing the performance of the iterative projection algorithm with that of INDTREES would be highly desirable—unfortunately, INDTREES is not available as an executable program. Hence, the INDTREES algorithm was reconstructed based on the original publication by Carroll et al. (1984) in the *Journal of Classification* and reprogrammed in `R` so that the performance of both algorithms could be illustrated by analyzing a real-world data set. (The `R` code is available from the authors upon request.)

2 Definitions and Concepts

2.1 Additive Trees and the Four-Point Condition

An additive tree is a weighted acyclic connected graph. The terminal nodes of an additive tree represent a set of N objects $\mathcal{O} = \{O_1, \ldots, O_N\}$. The weights along the paths connecting objects O_i, O_j, with $1 \leq i, j \leq N$—typically with a distance interpretation—can be collected into an $N \times N$ matrix $\boldsymbol{\Delta} = \{\delta_{ij}\}$. As a necessary and sufficient condition for a unique additive tree representation of $\boldsymbol{\Delta}$, the δ_{ij} must satisfy the additive inequality or four-point condition (Barthélemy & Guénoche, 1991; De Soete & Carroll, 1996; Semple & Steel, 2003):

$$\delta_{ij} + \delta_{kl} \leq \max\left\{\delta_{ik} + \delta_{jl}, \delta_{jk} + \delta_{il}\right\} \quad \text{for } 1 \leq i, j, k, l \leq N \tag{1}$$

or equivalently, for any object quadruple O_i, O_j, O_k, and O_l, the two largest sums of path length distances $\delta_{ij} + \delta_{kl}$, $\delta_{ik} + \delta_{jl}$, and $\delta_{jk} + \delta_{il}$ must be equal. The additive inequality is a generalization of the ultrametric inequality $\delta_{ij} \leq \max\left\{\delta_{ik}, \delta_{jk}\right\}$ for $1 \leq i, j, k \leq N$. Equivalently, an ultrametric defined on $\boldsymbol{\Delta} = \{\delta_{ij}\}$ for any object triple O_i, O_j, and O_k, requires that the largest two path length distances among δ_{ij}, δ_{ik}, and δ_{jk} be equal.

2.2 The Ultrametric-Star-Tree Decomposition of Additive Trees

Carroll (1976) discusses the decomposition of an additive tree distance into the sum of an ultrametric and a centroid distance. The latter induces a star as its tree representation, where one of the vertices, the center v_c, is fixed. Distances for all v_j, $v_{j'}$, with $j, j' \neq c$, are obtained by passing through v_c. Thus, any additive tree matrix $\boldsymbol{\Delta}$ can be decomposed into an ultrametric matrix $\mathbf{U} = \{u_{ij}\}$ and a centroid matrix $\mathbf{C} = \{c_{ij}\}$.

2.3 Constructing Additive Trees as a Constrained Least-Squares Problem

Let $\mathbf{P} = \{p_{ij}\}$ denote an $N \times N$ square-symmetric (two-way one-mode) proximity matrix. Finding an additive tree representation of $\mathbf{P}$ requires the estimation of path-length distances $\hat{\delta}_{ij} = d_{ij}$ that minimize the constrained least-squares loss function (see Barthélemy & Guénoche, 1991; De Soete & Carroll, 1996):

$$\min_{\mathbf{D}} \{L(\mathbf{D})\} = \min_{\mathbf{D}} \left\{ \sum_{i<j} (p_{ij} - d_{ij})^2 \right\} = \min_{\mathbf{D}} \left\{ \frac{1}{2} \mathrm{tr}(\mathbf{P} - \mathbf{D})(\mathbf{P} - \mathbf{D})' \right\} \qquad (2)$$

subject to

$$d_{ij} + d_{kl} \leq \max \left\{ d_{ik} + d_{jl}, d_{jk} + d_{il} \right\} \quad \text{for } 1 \leq i, j, k, l \leq N$$

(Note that $\mathbf{D} = \{d_{ij}\}$.) Conceptually, constructing an additive tree for a given proximity matrix is a constrained optimization problem that requires (a) to determine a particular topology, or branching pattern of the additive tree by establishing constraints based on the additive inequality that most faithfully reflect the relations among the $\binom{N}{4}$ quadruples of given proximities; (b) to estimate tree or path-length distances such that $L(\mathbf{D})$ is minimized subject to the specific constraints defining the additive tree topology as identified in (a). Křivánek (1986) showed that constructing an additive tree is NP-hard. Thus, currently available algorithms for constructing additive trees are heuristics, with no guarantee of obtaining a globally-optimal solution.

2.4 Constructing Multiple Tree Structures

Carroll and Pruzansky (1980; see also, Carroll, 1976; De Soete & Carroll, 1996) proposed constructing multiple additive tree structures to a given proximity matrix by means of successive residualization. As an example, consider representing $\mathbf{P}$ by two additive trees (called a “bi-additive” tree structure): an initial additive tree $\mathbf{D}^{(1)}$ is constructed for $\mathbf{P}$; then, a second structure $\mathbf{D}^{(2)}$ is constructed for the residual matrix $\mathbf{P} - \mathbf{D}^{(1)}$. In an attempt to further improve the fit of the resulting bi-additive tree structure, the residuals $\mathbf{P} - \mathbf{D}^{(1)} - \mathbf{D}^{(2)}$ are added back to $\mathbf{D}^{(1)}$, followed by (re-)constructing $\mathbf{D}^{(1)}$, potentially better fitting $\mathbf{P} - \mathbf{D}^{(2)}$, thereby producing a revised residual matrix $(\mathbf{P} - \mathbf{D}^{(2)}) - \mathbf{D}^{(1)}$, and so on. The process continues by repetitively fitting the residuals from the second additive tree by the first, and the residuals from the first additive tree by the second, until the sequence converges.

3 The Algorithms: INDTREES and Iterative Projection

3.1 The INDTREES Algorithm

Carroll and Pruzansky (1980) proposed a mathematical programming algorithm for constructing additive trees that uses the ultrametric-star-tree decomposition as a vehicle for constructing an additive tree. INDTREES (Carroll et al., 1984) is a generalization of Carroll and Pruzansky (1980) algorithm to three-way data. Let $\mathbf{P}_s$, $s = 1, 2, \ldots, S$, denote proximity matrices observed on S data sources. INDTREES is initialized by identifying an ultrametric representation of the average of these individual proximity matrices through hierarchical clustering. The topology of this average ultrametric tree is then used for fitting the individual proximity matrices. The same-topology condition is imposed through a penalty term enforcing that for each object triple O_i, O_j, O_k, the same two pairs of distances be the largest two across all sources. For each source, conformity of the distance estimates to the ultrametric inequality—the two largest distances must be equal—is secured by a second penalty term. The source-specific additive trees are constructed in fitting the residual proximities $p_{ij(s)} - u_{ij(s)}$ by individual star components through least squares. INDTREES iterates through these estimation steps until convergence; the (nonlinear) conjugate gradient method by Fletcher and Reeves (1964) is used for minimizing the least-squares loss function. INDTREES allows for fitting multiple tree structures through the successive residualization of the input proximity data.

3.2 The Iterative Projection Algorithm

The iterative projection (IP) algorithm for constructing additive tree structures by Hubert and Arabie (1995) is an adaptation of Dykstra’s (1983) general IP algorithm for solving least squares minimization problems, with side constraints C representing a closed convex set: given a vector $\mathbf{x}$, find the best approximating vector $\mathbf{x}^* \in C$. In theory, $\mathbf{x}^*$ could be found directly by projecting $\mathbf{x}$ onto C (written as $\mathbf{x}^* = P_C(\mathbf{x})$).

In practice, however, this may pose extreme computational demands. As a solution, Dykstra proposed to re-express the constraints as $C = \bigcap_1^M C_m \neq \emptyset$, which allows for the decomposition of the (presumably) difficult computation $\mathbf{x}^* = P_C(\mathbf{x})$ into the easier task of iteratively projecting $\mathbf{x}$ onto the M closed convex sets of constraints, $C_1, \ldots, C_M$, thereby constructing a sequence $\mathbf{x}_t$, with $t = 0, 1, 2, \ldots$ that was proven by Boyle and Dykstra (1985) to converge to $\mathbf{x}^* \in C$. The sequence $\mathbf{x}_t$ is initialized by setting $\mathbf{x}_0 = \mathbf{x}$, followed by the projection of $\mathbf{x}_0$ onto C_1, resulting in $\mathbf{x}_1$ that in turn is projected onto C_2, producing $\mathbf{x}_2$ to be projected onto C_3, and so on. The difference between consecutive projections $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$ is often called the increment, or residual. The algorithm concludes its first cycle of projections onto sets $C_1, \ldots, C_M$, with the projection of $\mathbf{x}_{M-1}$ onto C_M producing $\mathbf{x}_M$, the input vector for the second projection cycle through $C_1, \ldots, C_M$. To guarantee convergence to $\mathbf{x}^*$, from the second cycle on, each time $C_1, \ldots, C_M$ are revisited in subsequent cycles, the increment from the previous cycle associated with that particular set must be removed from the vector before actually proceeding with the projection.

An additive tree is determined by the collection of constraints, $C_1, \ldots, C_M$, as defined by the four-point condition. Each C_m is associated with one of the $M = \binom{N}{4}$ object quadruples, given N objects. For a specific quadruple m of objects O_i, O_j, O_k, and O_l, the four-point condition translates into three possible inequality constraints, one of which must be satisfied by the six distances involved

$$\begin{aligned}
\delta_{ij} + \delta_{kl} &\leq \delta_{ik} + \delta_{jl} = \delta_{jk} + \delta_{il} \\
\delta_{ik} + \delta_{jl} &\leq \delta_{ij} + \delta_{kl} = \delta_{il} + \delta_{jk} \\
\delta_{il} + \delta_{jk} &\leq \delta_{ij} + \delta_{kl} = \delta_{ik} + \delta_{jl}
\end{aligned}$$

These three constraints form the set C_m. Each constraint can be expanded into four inequalities:

$$\begin{aligned}
&\delta_{ij} + \delta_{kl} \leq \delta_{ik} + \delta_{jl} && \Leftrightarrow \delta_{ij} + \delta_{kl} - (\delta_{ik} + \delta_{jl}) \leq 0 \\
&\delta_{ij} + \delta_{kl} \leq \delta_{jk} + \delta_{il} && \Leftrightarrow \delta_{ij} + \delta_{kl} - (\delta_{jk} + \delta_{il}) \leq 0 \\
&\left.\begin{aligned} \delta_{ik} + \delta_{jl} \leq \delta_{jk} + \delta_{il} \\ \delta_{jk} + \delta_{il} \leq \delta_{ik} + \delta_{jl} \end{aligned}\right\} && \Leftrightarrow \delta_{ik} + \delta_{jl} - (\delta_{jk} + \delta_{il}) = 0
\end{aligned}$$

The constraints $C_1, C_2, \ldots, C_M$ determine the topology of an additive tree. The tree distances $\mathbf{d}^* \in C = \bigcap_1^M C_m$ are estimated such that the least squares loss function $(\mathbf{p} - \mathbf{d})'(\mathbf{p} - \mathbf{d})$ is minimized ($\mathbf{p}$ and $\mathbf{d}$ denote vectorizations of the matrices $\mathbf{P}$ and $\mathbf{D}$, respectively). After initializing $\mathbf{d}_0 = \mathbf{p}$, the algorithm proceeds by checking for each quadruple of objects whether the involved distances conform to the respective constraints in C_m.

If a violation is encountered, the vector of distances is projected onto C_m, and the particular distances are replaced by their projections (Dykstra, 1983; Han, 1988).

The algorithm cycles through $C_1, \ldots, C_M$ until convergence. The fit of an additive tree structure is quantified by the variance-accounted-for criterion (VAF):

$$\mathrm{VAF} = 1 - \frac{\sum_{i<j}(p_{ij} - d^*_{ij})^2}{\sum_{i<j}(p_{ij} - \bar{p})^2}$$

with $\bar{p}$ denoting the mean of the off-diagonal entries in $\mathbf{P} = \{p_{ij}\}$.

4 Application: Judgments of Schematic Face Stimuli

A total of 22 graduate students in the Psychology department of the University of Illinois provided pairwise dissimilarity ratings of 12 schematic faces. The twelve face stimuli were generated by completely crossing the three factors "Facial Shape","Eyes", and "Mouth" (see Fig. 1).

Data Analysis: Part 1. INDTREES and the iterative projection algorithm share the rationale to model individual differences as deviations from a common frame of reference. Hence, source-specific, individualized additive trees are all restricted to have the same topology; individual variation is modelled through differential shrinking or stretching of the tree branch lengths. The VAF criterion obtained for each source serves as a fit index quantifying how closely the individual additive trees reflect the properties of the shared topology. Remarkably, the results obtained for INDTREES and the iterative projection algorithm were virtually indistinguishable. Hence, due to space restrictions, only the results of the iterative projection algorithm are reported.

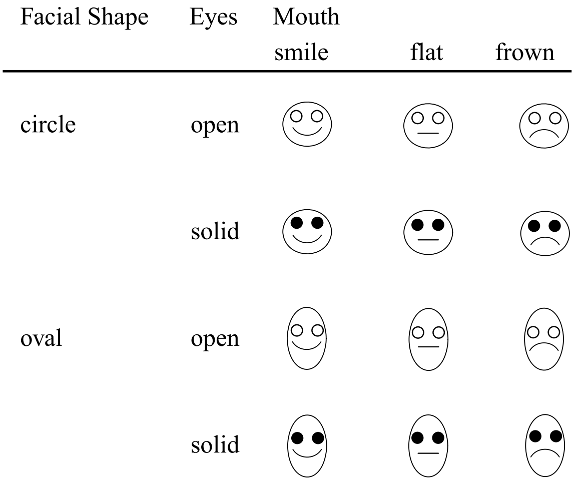


Fig. 1 The construction of schematic face stimuli

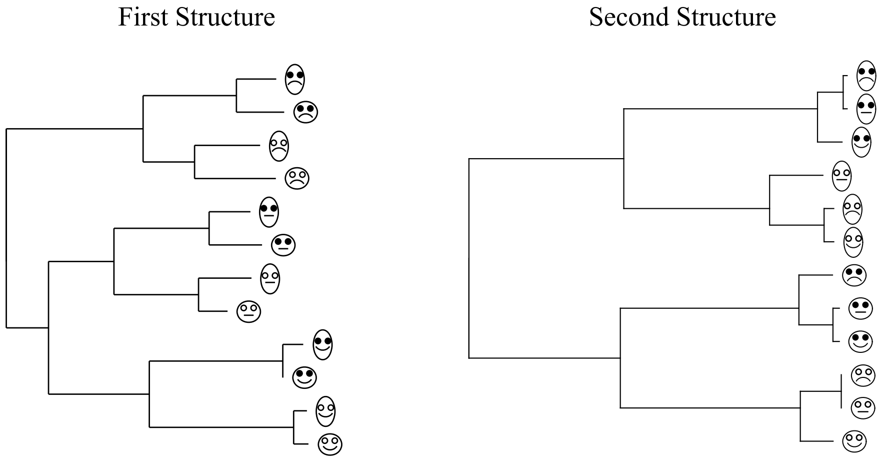


Fig. 2 Biadditive reference tree representation (VAF = .8521)

Results: Part 1. A bi-additive tree was chosen as reference structure for the 22 individual proximity matrices; the two trees are shown in Fig. 2. The first additive tree structure identifies three segments of schematic faces based on the primary criterion "emotional impression", as implied by the factor "mouth" with its levels"frown", "flat", and"smile". Apparently, the three categories are not perceived as equally distinct: "flat" and"smile" are merged, whereas "frown" is set apart. Within each group, a secondary distinction into faces with "open" versus"solid" circled eyes can be observed; the factor "facial shape" distinguishes between stimuli at a tertiary level. The second additive tree structure produces a perfectly balanced grouping dominated by "facial shape"; the contrast between "open" and"solid" circled eyes appears to be a secondary criterion, whereas emotional impression serves as tertiary criterion.

For each of the 22 subjects, the VAF score was computed indicating how well each subject's data were actually represented by the bi-additive reference structure. The 22 subjects were ranked according to their VAF scores. Based on this ranking, the top-three subjects, who were especially well-represented by the bi-additive reference structure, and the bottom-three (i.e., worst-represented) subjects were collected into two extreme groups that formed the focus of the subsequent analysis. (Comparing these two extreme groups mirrored a paradigm in experimental psychology to emphasize the effect of a particular treatment under study.) The top-three group consisted of Subjects 5, 12, and 17; the bottom-three group consisted of Subjects 1, 14, and 2. From the top-three group, only Subject 5 is further considered here because her results are almost identical to those of Subjects 12 and 17. However, Subjects 1, 14, and 2 are all considered here because the slight variations in the additive tree representations of their data are instructive for further understanding how the face stimuli were perceived. The additive tree graphs of Subjects 5, 1, 14, and 2 are presented below, split into two displays, Figs. 3 and 4. Due to space limitations, only the first additive tree structures are shown.

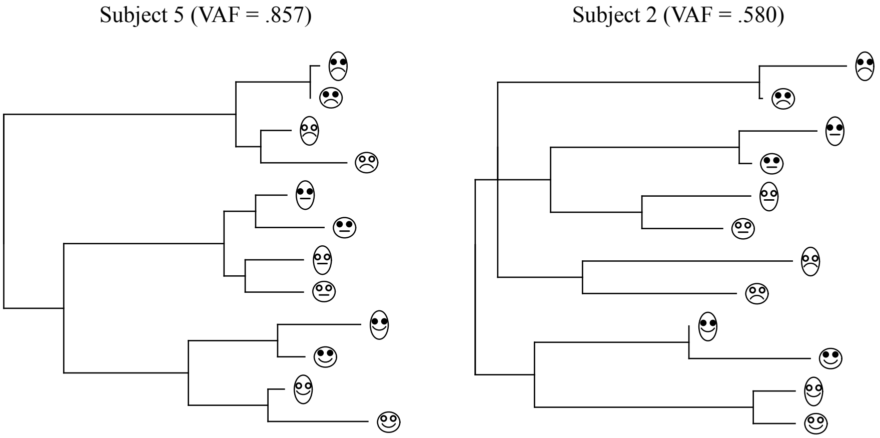


Fig. 3 Individual biadditive tree representations for selected subjects: first structure

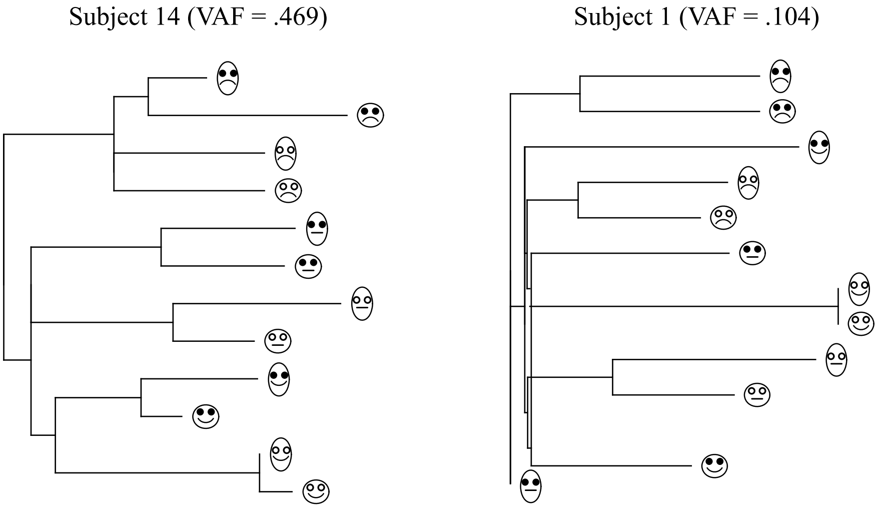


Fig. 4 Individual biadditive tree representations for selected subjects: first structure

Not surprising, Fig. 3 confirms that the tree of Subject 5 almost perfectly matches the reference structure. Note that the tree structure of Subject 2 displays several ties between face stimuli—indicated by horizontal bars joining more than two vertical branches at a time. Ties can occur as a result of remedying violations of the four-point condition by averaging the involved distances. Specifically, for Subject 2, the "smile" segment conforms to the reference structure, whereas the "flat" and"frown" categories are mingled such that the "frown" category is draped around the"flat" segment.

As Fig. 4 shows, similar observations can be made about the additive tree of Subject 14: the faces in the “flat”,“smile”, and “frown” segments are not well separated—apparently, conformity to the reference structure could only be enforced through many tied distance estimates. Finally, the individual additive tree structure obtained for Subject 1 is extremely distorted with its numerous ties; this tree is essentially non-interpretable.

Data Analysis: Part 2. To further explore the seemingly deviant perceptions of the face stimuli by Subjects 2, 14, and 1, their dissimilarity data were re-analyzed separately, without using the bi-additive reference structure from the previous analysis step. Such an independent re-analysis of the data was supposed to clarify the question whether the lack of fit observed with Subjects 2, 14, and 1 was due to (erratic) random responses, or indicated a coherently different view on the face stimuli, driven by idiosyncratic criteria that simply did not match the rest of the sample. Each individual proximity matrix was fitted by a bi-additive tree structure.

Results: Part 2. Due to space restrictions, displays only of the first additive tree structures for Subjects 1 and 14 are presented here (see Fig. 5). The arrangement of the face stimuli in the tree graph for Subject 1 does not reveal a discernable pattern, which suggests that Subject 1 provided random judgments—a likely explanation, given that subjects were required to participate in this study for credit. The tree diagram of Subject 14 tells a different story. Compared to the reference tree, Subject 14 assigned different priorities to the criteria for judging the face stimuli: “eye shape” is the primary criterion to distinguish between the twelve schematic faces, whereas “facial shape” served as secondary criterion.

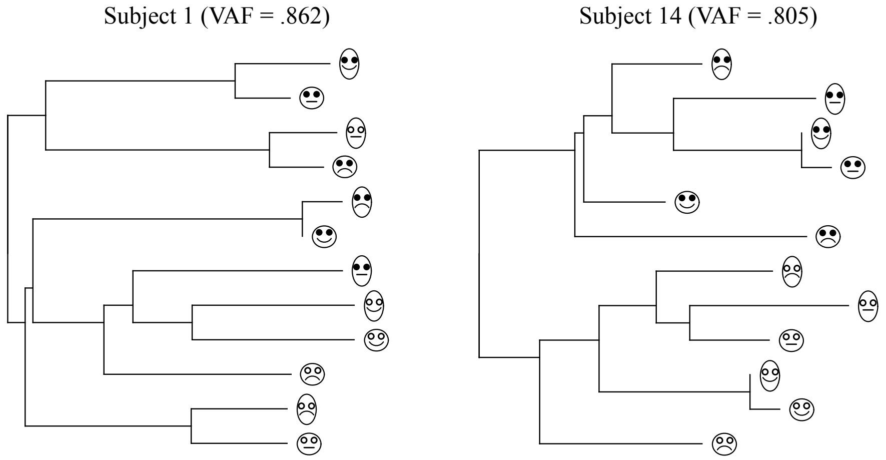


Fig. 5 Independent biadditive tree representations for subjects 1 and 14: first structure

5 Conclusion

A comprehensive evaluation of three additive tree fitting algorithms by Smith (1998) showed that IP consistently outperformed ADDTREE (Corter, 1982; see also Sattath & Tversky, 1977) and De Soete's (1983) penalty function algorithm in finding the best-fitting additive tree. The study presented in this paper appears to be the first to compare the performance of IP in fitting additive tree structures to three-way data with that of INDTREES, apparently the only extant penalty function algorithm for fitting additive trees to three-way data. Both algorithms—although perhaps not immediately obvious—rely on a principle common in statistics as well as of immediate intuitive appeal, namely, to analyze individual variability against an average reference structure. For the illustrative example chosen here, the results of the two algorithms were indistinguishable. However, further research is needed to assess their performance with a broader scope.

An another question is how the reference-structure approach might compare to consensus tree modeling, an alternative way to analyze three-way data rooted in a different philosophy. Initially, separate tree representations are fitted to each individual data source, followed by an integrative step aimed at locating a prototypical tree representation that captures a maximum of the individual variability observed among the independent source trees, analogous to a majority voting rule.

As a final note, additive tree representations appear to have been under-used in psychology as a data analytic tool. The reason might be the relative inaccessibility of suitable software in the past. The situation is different today: the potential user can choose among a variety of programs and software packages. For example, an implementation of the ADDTREE algorithm is available in `SYSTAT`; Hubert et al. (2006) provided `MATLAB` routines for fitting additive trees that have recently also been made available in the `R` package `clue` (Hornik, 2018).

References

Barthélemy, J. P., & Guénoche, A. (1991). *Tree and proximity representations*. Chichester: Wiley.

Boyle, J. P., & Dykstra, R. L. (1985). A method for finding projections onto the intersection of convex sets in Hilbert spaces. In R. L. Dykstra, R. Robertson, & F. T. Wright (Eds.), *Advances in order restricted inference*, Lecture Notes in Statistics (Vol. 37, pp. 28–47). Berlin: Springer.

Carroll, J. D. (1976). Spatial, non-spatial and hybrid models for scaling. *Psychometrika*, *41*, 439–463.

Carroll, J. D., & Pruzansky, S. (1980). Discrete and hybrid scaling models. In E. Lantermann & H. Feger (Eds.), *Similarity and choice* (pp. 108–139). Bern: Huber.

Carroll, J. D., Clark, L. A., & DeSarbo, W. S. (1984). The representation of three-way proximities data by single and multiple tree structure models. *Journal of Classification*, *1*, 25–74.

Corter, J. E. (1982). ADDTREE/P: A PASCAL program for fitting additive trees based on Sattath and Tversky's ADDTREE algorithm. *Behavior Research Methods and Instrumentation*, *14*, 353–354.

De Soete, G. (1983). A least-squares algorithm for fitting additive trees to proximity data. *Psychometrika*, *48*, 621–626.

De Soete, G., & Carroll, J. D. (1989). Ultrametric tree representations of three-way three-mode data. In R. Coppi & S. Belasco (Eds.), *Analysis of multiway data matrices* (pp. 415–426). Amsterdam: North Holland.

De Soete, G., & Carroll, J. D. (1996). Tree and other network models for representing proximity data. In P. Arabie, L. J. Hubert, & G. De Soete (Eds.), *Clustering and classification* (pp. 157–197). River Edge, NJ: World Scientific.

Deutsch, F. (2001). *Best approximation in inner product spaces*. New York: Springer.

Dykstra, R. L. (1983). An algorithm for restricted least-squares regression. *Journal of the American Statistical Association*, *78*, 837–842.

Fletcher, R., & Reeves, C. M. (1964). Function minimization by conjugate gradients. *Computer Journal*, *7*, 149–154.

Han, S. P. (1988). A successive projection method. *Mathematical Programming*, *40*, 1–14.

Hubert, L. J., & Arabie, P. (1995). Iterative projection strategies for the least-squares fitting of tree structures to proximity data. *British Journal of Mathematical and Statistical Psychology*, *48*, 281–317.

Hubert, L. J., Arabie, P., & Meulman, J. (2006). *The structural representation of proximity matrices with MATLAB*. Philadelphia, PA: SIAM.

Hornik, K. (2018). *clue: Cluster ensembles. R package version 0.3-56*. Retrieved from the Comprehensive R Archive Network [CRAN] website https://cran.r-project.org/web/packages/clue/

Křivánek, M. (1986). On the computational complexity of clustering. In E. Diday, Y. Escouffier, L. Lebart, J. P. Pagés, Y. Schektman, & R. Tomassone (Eds.), *Data analysis and information, IV* (pp. 89–96). Amsterdam: North Holland.

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika*, *42*, 319–345.

Semple, C., & Steel, M. (2003). *Phylogenetics*. Oxford, UK: Oxford University Press.

Smith, T. J. (1998). A comparison of three additive tree algorithms that rely on a least-squares loss criterion. *The British Journal of Mathematical and Statistical Psychology*, *51*, 269–288.

A Comparison of Ideal-Point and Dominance Response Processes with a Trust in Science Thurstone Scale

Samuel Wilgus and Justin Travis

Abstract The purpose of this study is to compare the dominance and ideal-point response process models for a trust in science measure developed from Thurstone's (Am J Sociol 33(4):529–554, 1928; Psychol Rev 36(3):222–241, 1929) scaling procedures. The trust in science scale was scored in four different ways: (1) a dominance response approach using observed scores, (2) a dominance response approach using model-based trait estimates, (3) an ideal-point response observed score approach using Thurstone scoring, and (4) an ideal-point response approach using model-based trait estimates. Comparisons were made between the four approaches in terms of psychometric properties and correlations with political beliefs, education level, and beliefs about scientific consensus in a convenience sample of 401 adults. Results suggest that both the ideal-point and two-parameter IRT models fit equally well in terms of overall model fit. However, two items demonstrated poor item fit in the two-parameter model. Correlations with political beliefs, education level, and science-related items revealed very little differences in magnitude across the four scoring procedures. This study shows support for the flexibility of the ideal-point IRT model for capturing non-ideal-point response patterns. The study also demonstrates the use of using IRT to examine item parameters and item fit.

Keywords Dominance response process · Ideal-point response process · Thurstone scaling

S. Wilgus (✉)
North Carolina State University, 2310 Stinson Dr., Raleigh, NC 27695, USA
e-mail: sjwilgus@ncsu.edu

J. Travis
University of South Carolina Upstate, Spartanburg, SC 29303, USA

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_36

1 Introduction

1.1 Dominance Response Process

Classical test theory (CTT) has provided the most common approach for both developing and scoring psychological measures. Likert (1932) and Thurstone (1928) were early pioneers on survey methodologies, although Likert's graded response (agree-disagree continuum) has persisted as the relative "gold standard" for constructing psychological measures. Specifically, Likert's (1932) approach involves selecting items that show high item-to-total score correlations so that the measure demonstrates high internal consistency and exhibits a single-factor solution. Importantly, traditional approaches underlying the vast majority of Likert scales available in psychometric literature assume a monotonic relationship between responses across item-s—a dominance model. The dominance model proposes that high levels of a given attribute result in higher probabilities of agreement across items. This results in an item response function (IRF) that resembles a monotonically increasing S-shaped curve and an item information function that is maxed at a single level of the latent trait continuum.

Methods for scoring psychological scales using observed scores or model-assigned latent trait estimates (i.e., item response theory or IRT) scores can be used to reflect an individual's standing on a trait using the dominance response process. The vast majority of scoring procedures using observed scores reflect the dominance approach by simply reverse-scoring any reverse coded items and then taking a mean or sum of item responses. Higher scale scores reflect higher standing on the latent trait. Item response theory (IRT) models use characteristics of items, such as an item's difficulty, discrimination, and guessability to establish a link between an individual's response pattern and their standing on the latent continuum. IRT models such as the popular Rasch, two-parameter (2PL), and three-parameter (3PL) models, reflect the dominance response model by assuming that higher standing on the latent trait results in higher probabilities of endorsing items of the scale.

1.2 Ideal-Point Response Process

Although some attributes (e.g., cognitive ability) may be best measured via dominance approaches, other attributes (e.g., personality or attitudes) may be more appropriately described by an ideal-point response process. The ideal-point response process differs from the dominance approach with higher levels of an attribute not necessarily resulting in higher endorsement across items. Instead, endorsement is maxed around what later psychometricians would call an ideal-point, or a particular area along a latent trait's continuum that closely resembles the underlying level of the responder's attribute (Coombs, 1964). This results in an item response function that is bell-shaped, such that the probability of endorsing an item is maxed at the appropriate level of the latent trait. The probability of endorsing an item then decreases as an individual's latent trait level moves below *and* above the "ideal point" such that an individual can have a low probability of endorsing an item because their standing on the latent trait is "too high" *or* "too low." This results in an item information function that is bimodal because individuals can choose to not endorse an item due to having too low of standing on the latent trait *or* having too high of standing on the latent trait. For example, an item intended to measure conscientious such as "My room neatness is about average" would receive endorsements from individuals in the middle of the latent continuum. However, this item may also result in some individuals disagreeing with item because they have below average room neatness (disagreeing from below) and individuals disagreeing with the item because they have above average neatness (disagreeing from above).

Just as observed and model-based scoring approaches exist for the dominance approach, there are multiple ways to score psychological scales that reflect an ideal-point approach. Thurstone's (1928) method of scale development and scoring reflects a method of assigning scores that take into account individuals disagreeing from below *and* above by adding an extra step to the scale development process where subject matter experts assign ratings to each item from 1 (extremely low on the latent continuum) to *n* (*n* being the total number of items on the scale, reflecting the most extreme items on the latent continuum). Subject matter expert ratings are then averaged for each item to result in an expert-based trait score. Respondents are assigned the highest trait score across all of the items they endorsed. Thus, if a respondent disagreed with all items except for the most extreme item, which they endorsed, they would still receive the highest score possible because of the assumption that they disagreed from above with all of the less extreme items. An example of a Thurstone scale can be seen in Roberts and Laughlin's (1996) analysis of an older Thurstone scale on attitudes toward capital punishment. Consistent with Thurstone's (1928) theory of scaling, Roberts and Laughlin considered extreme items (e.g., "Capital punishment is never justified" or "Capital punishment should be used more than it is") and intermediate items (e.g., "I do not believe in capital punishment, but it is not practically advisable to abolish it").

Model-based approaches also exist for estimating an individual's standing on a latent trait under ideal-point assumptions. For example, the generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 2000) allows for the typical monotonic s-shaped IRF curve to "unfold," resulting in a bell-shaped IRF that reflects the probability of endorsing an item decreasing when either below *or* above the ideal-point, in contrast to assuming the monotonically increasing IRF of dominance approaches.

1.3 Dominance Versus Ideal-Point Response Process in Psychological Measurement

Despite the ability to capture a wide range of a latent continuum, Thurstone's method has not seen widespread use like Likert's method, likely due to the added step during scale development. This is unfortunate, given that that recent research has found that a dominance model may not always be appropriate, especially when it comes to personality and attitude data (Drasgow, Chernyshenko, & Stark, 2009; Roberts et al., 2000). This is especially the case for neutral items such as "My room neatness is about average" which are typically discarded from scales build under the dominance approach due to poor item-total correlations (likely due to individuals disagreeing with the item from below and above on the latent continuum). Indeed, most psychological scales only contain either positively- or negatively-worded items (e.g., "Being neat is not exactly my strength"), resulting in higher internal consistency but less representation of the full latent continuum.

Given that most of the research applying ideal-point analyses has used scales developed through dominance response process assumptions, very little in known as to how ideal-point analyses compare to dominance analyses for scales developed from ideal-point procedures, such as scales developed using Thurstone's method of scale development. Indeed, Chernyshenko, Stark, Drasgow, and Roberts (2007) mention the need for investigating scales built on the assumption of representing an unfolding response process from inception, such as Thurstone scales. This paper answers calls from the psychometric literature to investigate the utility of ideal-point model assessment of measures developed with ideal-point scaling methods (Chernyshenko et al., 2007; Drasgow et al., 2009; Stark, Cheryshenko, Dasgow, & Williams, 2006).

1.4 The Present Study

The purpose of the current study is to compare whether a dominance approach or an ideal-point approach is more appropriate for assessing an individual's trust in science, a measure developed using the Thurstone method (1928). Initial investigations have demonstrated that dominance models may be inappropriate for analyzing some attitude measures (Stark et al., 2006), yet there has been scant follow-up from psychometricians in psychological fields (see Drasgow, Chernyshenko, & Stark, 2010 for a discussion). The implications of our investigation are two-fold.

First, many constructs that are important to psychologists entail wide-ranging phenomena that may be more theoretically congruent with an ideal-point response process than a dominance response process. That is, survey items requesting a respondent's agreement with a statement are likely to require a process of introspection where proximity of the item to their actual, or latent, level of the attribute being measured. While this is implicit in the early psychometric approaches to survey response processes, subsequent cognitive approaches propose information processing mechanisms that necessitate identification of an item along a continuum and a comparison of one's self with that item's location (see Tourangeau, Rips, & Rasinski, 2000).

Second, psychological measures impact areas of life outside of the lab, such as personnel selection, training, clinical assessment, and many other applications. If scales are developed and analyzed under the guidance of dominance process assumptions while not meeting these assumptions, then both the science and practice of psychological measurement will be restricted in utility. Two areas that have recently made strides in the use of ideal-point analyses are personality assessment and performance evaluation. Specifically, forced-choice formats can be used to provide accurate measures of job performance (Borman, Penner, Allen, & Motowidlo, 2001) and personality (Chernyshenko et al., 2009). Ipsative properties of forced choice formats common in personnel psychology are problematic (see Meade, 2004 for discussion), and recent developments in ideal-point IRT analyses and computer adaptive testing have potential to overcome these limitations.

In sum, assumptions of a dominance response process may not be suitable for many common attitudinal surveys and an ideal-point process may be more favorable. Further, little research has revisited the Thurstone method of scale construction with ideal-point analyses, instead opting to apply these analyses to Likert-type scales already in use. The current study examines four analytic techniques (CTT-dominance, CTT-ideal-point, IRT-dominance, and IRT-ideal-point) to assess whether an ideal-point or dominance process fits observed data better.

1.5 Hypotheses

As mentioned previously, researchers have only recently begun applying ideal-point analyses but almost exclusively with established scales developed under dominance frameworks. As an exception, Chernyshenko and colleagues (2007) developed a measure of the conscientiousness facet order using an ideal-point process. These authors found their ideal-point IRT analyses to provide a substantial improvement in test information and measurement precision over dominance IRT and classical test theory analyses. Provided the aforementioned theoretical assertions and this recent empirical finding, we predict that ideal-point scoring of the trust in science scale will outperform the traditional dominance approach to scoring in terms of item fit and resulting relationships between trait estimates and convergent criteria.

Convergent criteria were selected on the basis of rational and empirical relationships with trust in science. These included, political beliefs, education level, and several statements regarding beliefs in scientific conclusions (e.g., vaccines cause autism). For example, a 2014 study from the Pew Research Center found that democrats and liberals were more likely to agree with scientific consensus about matters, such as climate change, than their conservative and republican counterparts (Funk & Rainie, 2015). Furthermore, educational attainment, certainly in science-based fields, may be somewhat related to trust in science as recent research has found that increased academic education has a positive effect on knowledge of climate change (McCright, 2010). Lastly, several statements about popular issues for which scientific consensus exists were chosen as they represented several statements reflecting beliefs that are (in)congruent with modern scientific findings.

2 Method

2.1 Sample

Our sample consisted of 401 respondents who were recruited to take the online survey through posting the survey on various social media platforms across the 21 doctoral students that contributed to item generation.

2.2 Measures

Trust in science. The Thurstone-type scale assessed in this study was developed to create an instrument that would measure the extent to which an individual trusted science. The scale consisted of nine statements that reflected attitudes towards science, scientific method, and applications of science. The statements were generated through 21 graduate students in psychology doctorate programs as part of a psychometrics course. The professor of the course who led the scale development effort specialized in psychometrics. In line with the procedures described by Thurstone (1928, 1932), the scale development began with generating a large bank of statements that reflected varying attitudes towards science. All statements were later evaluated by the graduate students and assigned a rating from 1 (people endorsing this item show the least amount of trust in science) to 9 (people endorsing this item show the most amount of trust in science). The final 9 items were chosen on the basis of (1) finding items that had ratings along the entire continuum of 1–9 and (2) the level of consensus based on the standard deviation in graduate student ratings for each item. The final nine statements were presented to survey respondents along with asking if the respondent (1) agreed or (0) disagreed with each statement.

Political beliefs. Political beliefs were assessed with a single item asked, “How would you describe your political views?” on a five-point scale with 1 = “Very liberal” to 5 = “Very conservative.”

Education level. Education level was assessed with a single item that asked, “What is the highest level of school you have completed or the highest degree you have achieved?” with options including 1 = high school, 2 = some college, 3 = associate’s degree, 4 = bachelor’s degree, 5 = master’s degree, 6 = professional degree (JD, MD), or doctoral degree.

Beliefs in Scientific Consensus. To further examine the convergent validity of the trust in science scale, six additional statements that were expected to show significant correlations with trust in science were written by the authors and presented to respondents. These statements reflected current issues for which there is scientific consensus, such as whether vaccines cause autism and statements about climate change (e.g., "Scientists do not agree on whether the planet is warming or not."). These items were scored on a scale of 1 = strongly disagree to 5 = strongly agree.

2.3 Scoring Procedure

The trust in science scale was scored in four different ways: (1) a dominance response approach using observed scores, (2) a dominance response approach using model-based trait estimates, (3) an ideal-point response observed score approach using Thurstone scoring, and (4) an ideal-point response approach using model-based trait estimates. The first scoring approach reflected the traditional CTT order scale approach to measurement by taking a summed score across the 9 statements. Items that were negatively worded (e.g., "Science is a hoax") were reverse scored such that all the item-total correlations were positive. A total score was computed through summing across all items despite low internal consistency. The second scoring approach used the 2PL IRT model which estimates latent trait levels based on response patterns, item difficulty, and item discrimination. The third scoring approach was based on the scoring procedure described by Thurstone (1928). As mentioned earlier Thurstone scaling approach assigns the mean rating from the graduate student ratings when a respondent agrees with the statement. For example, if a respondent agrees to with the statement "Only knowledge gained through the scientific method can be trusted," they receive the mean rating of 8.45 for that endorsement (Table 1).

The respondent's final score was taken from their highest mean rating across the items. The fourth scoring method was conducted using the dichotomous ideal-point

Table 1 Trust in science scale with mean and standard deviations from subject matter expert ratings

Item		Mean	SD
1	Science is a hoax	1.15	.45
2	Scientific research does more harm than good	2.05	.60
3	Personal experience is more important than scientific findings	3.1	.72
4	Not all research is trustworthy	4.25	.72
5	Science is one source of knowledge among many	4.9	.37
6	Science can sometimes be helpful when making decisions	6	.65
7	Most research can be trusted	7.05	.76
8	Science is our best chance of solving the world's problems	8.25	.85
9	Only knowledge gained through the scientific method can be trusted	8.45	.89

IRT model (Maydeu-Olivares, Hernandez, & McDonald, 2006) as part of the 'mirt' package for IRT in R to generate trait estimates after fitting the data to an ideal-point model.

2.4 Fit Analyses

Our first hypothesis predicted that the ideal-point model would fit the data better than the two-parameter model. Before comparing models, two key assumptions must be examined to have confidence in the parameters obtained from an IRT model: unidimensionality and local independence. Whether these assumptions are met can be examined by assessing item fit indices. Before examining these assumptions, three items were dropped due to extreme response rates creating difficulty in estimating fit ("Science is a hoax", "Science can sometimes be helpful when making decisions", and "Not all research is trustworthy"). These three items had less than 2.5% of the sample either agreeing or disagreeing. After discarding these three items, conformability to the IRT assumptions were examined using the S-X^2 statistic to evaluate the extent to which observed data matched what was expected by the model (Orlando & Thissen, 2003). Within the ideal-point model, one item ("Scientific research does more harm than good") demonstrated poor fit (S-X$^2 = 6.38, p < .05$) and was removed.

3 Results

Descriptive statistics for all trust in science scale items, political beliefs, education levels, and beliefs in scientific consensus items are listed in Table 2. Respondents' average political belief rating was 2.31, which is between 2 (liberal) and 3 (slightly liberal). In terms of education attainment, 41.4% of participants listed a high school degree as the highest education level attained, 42.64% indicated some college, and 15.96% reported having earned a graduate or professional degree.

3.1 Examining Model Fit

As seen in Table 3, all S-X^2 values for the ideal-point model indicate all items show adequate fit to the model. However, two items showed poor fit to the two-parameter model. Despite the two poor-fitting items in the two-parameter model, additional fit indices supported the data fitting both models overall, with the ideal-point M2 = 6.47, p = .26; RMSEA = .027; TLI = .97; CFI = .99 and the two-parameter model M2 = 3.72, p = .59; RMSEA > .005; TLI = 1.02, CFI = 1.00. Thus, our first hypothesis that predicted the ideal-point model would fit the data better than the two-parameter model was only partially supported with the two-parameter fitting

Table 2 Means and standard deviations for trust in science scale items, political beliefs, education level, and beliefs about scientific consensus

Item	Mean	SD
Scientific research does more harm than good	.08	.27
Most research can be trusted	.70	.46
Science is one source of knowledge among many	.90	.30
Science is a hoax	.01	.10
Only knowledge gained through the scientific method can be trusted	.17	.37
Science can sometimes be helpful when making decisions	.99	.10
Not all research is trustworthy	.98	.15
Personal experience is more important than scientific findings	.19	.39
Science is our best chance of solving the world's problems	.64	.39
Political beliefs	2.31	1.11
Education level	2.68	2.74
I believe non-traditional approaches to medicine can be just as effective as those based on scientific studies	3.01	1.26
You can't trust information from the main stream media	3.42	1.20
Scientists do not agree on whether the planet is warming or not	2.06	1.31
Global warming is real, and it is caused by humans	4.23	1.04
I believe in the power of prayer to heal the body	2.82	1.51
Vaccines can cause autism	1.66	1.02

Table 3 Item parameters and fit statistics across the ideal-point model and two-parameter model

Item	Ideal-point model			Two-parameter model		
	a	b	S-X^2	a	b	S-X^2
Most research can be trusted	1.30	5.01	1.18	1.34	−1.43	.12
Science is one source of knowledge among many	2.36	2.49	.42	1.96	1.78	2.59
Only knowledge gained through the scientific method can be trusted	.14	21.32	1.22	.28	−3.02	8.07**
Personal experience is more important than scientific findings	2.25	−3.01	.16	2.12	−.46	4.38*
Science is our best chance of solving the world's problems	1.51	−5.00	1.29	1.61	1.42	.08

*p < .05, **p < .01

equally well, if not somewhat better overall, but demonstrating poor individual item fit for two of the five items.

Table 4 Means, standard deviations, and bivariate correlations for scores generated across the different scoring procedures

Scoring procedure		Mean	SD	1	2	3
1	Ideal-point model	.00	.73	–		
2	Two-parameter model	.00	.76	.98**	–	
3	CTT scaling	5.53	1.22	.89**	.88**	–
4	Thurstone scaling	7.74	.85	.79**	.79**	.73**

***p* < .01

3.2 Comparison of Scoring Procedures

Our second hypothesis predicted that the ideal-point model theta estimates (i.e., latent trait estimates) and Thurstone-scaled scores would demonstrate stronger relationships with items that should be related to trust in science. To test this hypothesis, bivariate correlations were examined between political affiliation, education, and six statements about trust in science (Table 4). Significant correlations were found between all four scoring procedures and political affiliation, with higher conservativism being associated with a lower trust in science. In addition, significant correlations were found between all four scoring procedures and all of the related items about trust in science. No significant correlations were found between the four score sets and education level. As indicated in Table 4, very similar correlation coefficients were found between the four scoring procedures and each item, with the exception of "I believe in the power of prayer to heal the body" which both IRT models showed notably larger correlations with than the classical test theory and Thustone scaling scores. The few differences in magnitude across the four scoring methods did not seem to have any other pattern. Thus, our second hypothesis was largely not supported. Examining the intercorrelations between the four sets of scores revealed that the theta estimates from the two IRT models and the CTT-scaled scores had very high similarity while the Thurstone scaling was less similar to the other three sets of scores, although still strongly related (Table 4).

4 Discussion

This study aimed to examine a scale through ideal-point approaches and dominance approaches. Within the ideal-point and dominance approaches, this study also examined differences in using model-based measurement via ideal-point and two-parameter IRT models and observed score-based measurement via classical test theory scaling and Thurstone scaling. Results indicated that although the two-parameter model fit just as well as the ideal-point model overall, two out of five items showed poor fit to the two-parameter model. Across the four scoring procedures, very little difference was found in relationships with political affiliation, education, and six statements about trust in science. The similarities in relationships with other variables as well as both models fitting the data overall are better explained when examining the item characteristic curves across both models (Fig. 1a, b).

Although the ideal-point model indicated a peak where the probability of endorsing the item would be maxed, this point was either far above or below reasonably common theta levels. This resulted in an item response function that appeared similar to a two-parameter model for actual *observed* levels of theta (Fig. 2). Despite this, the fact that some individual items showed better fit to the ideal-point model may suggest that the ideal-point model is more flexible at modeling both unfolding items as well as dominance items, which is consistent with past conclusions from similar research (Chernyshenko et al., 2007).

This study also suggests that, despite creating a scale through ideal-point assumptions, item characteristic curves consistent with the dominance assumptions of item responding are still found. However, the Thurstone scaling, which represents ideal-point assumptions, was found to correlate less with the other scoring procedures. In addition, the Thurstone scaling explained unique variance with the other items despite showing correlations with the other items similar in magnitude to the other scoring procedures. This may suggest that the trust in science scale items do possess some degree of unfolding that is not captured in the ideal-point IRT model.

However, the lack of unfolding occurring across levels of theta estimated from the sample may be related to this study's limitations. First, the sample that was collected consisted of friends, family, and colleagues of 21 doctoral students. This approach of sampling was not random and likely is biased to be higher educated and familiar with science than a typical random sample. As a result, the range of theta examined may not capture levels of theta that exist in the rest of the population. Theta levels not captured in the present sample may exist at points where unfolding may occur for some items. In addition, discarding some of the initial nine items due to extreme ratios of agree/disagree was also likely due to sample characteristics. Only examining a final five items due to item removals prevented the number of unique theta levels that could be estimated, further removing opportunities for any possible unfolding to be captured.

Future research should address the limitations this study had to better examine comparisons between ideal-point approaches and dominance approaches. Namely, future research should administer scales developed from an ideal-point perspective, such as Thurstone scales, to a large random sample where a wider range of latent trait levels can be observed. In addition, future studies should attempt to begin with larger banks of items in preparation for reducing the number of items due to poor fit, such as in this study and prior research examining IRT models (Chernyshenko et al., 2007; Roberts & Laughlin, 1996; Stark et al., 2006).

This study represented the first examination of a scale built from ideal-point assumptions across several scoring procedures. In addition to providing additional comparison across these scoring procedures, this study further demonstrates the advantage of IRT in being able to examine item parameters and item-data fit to

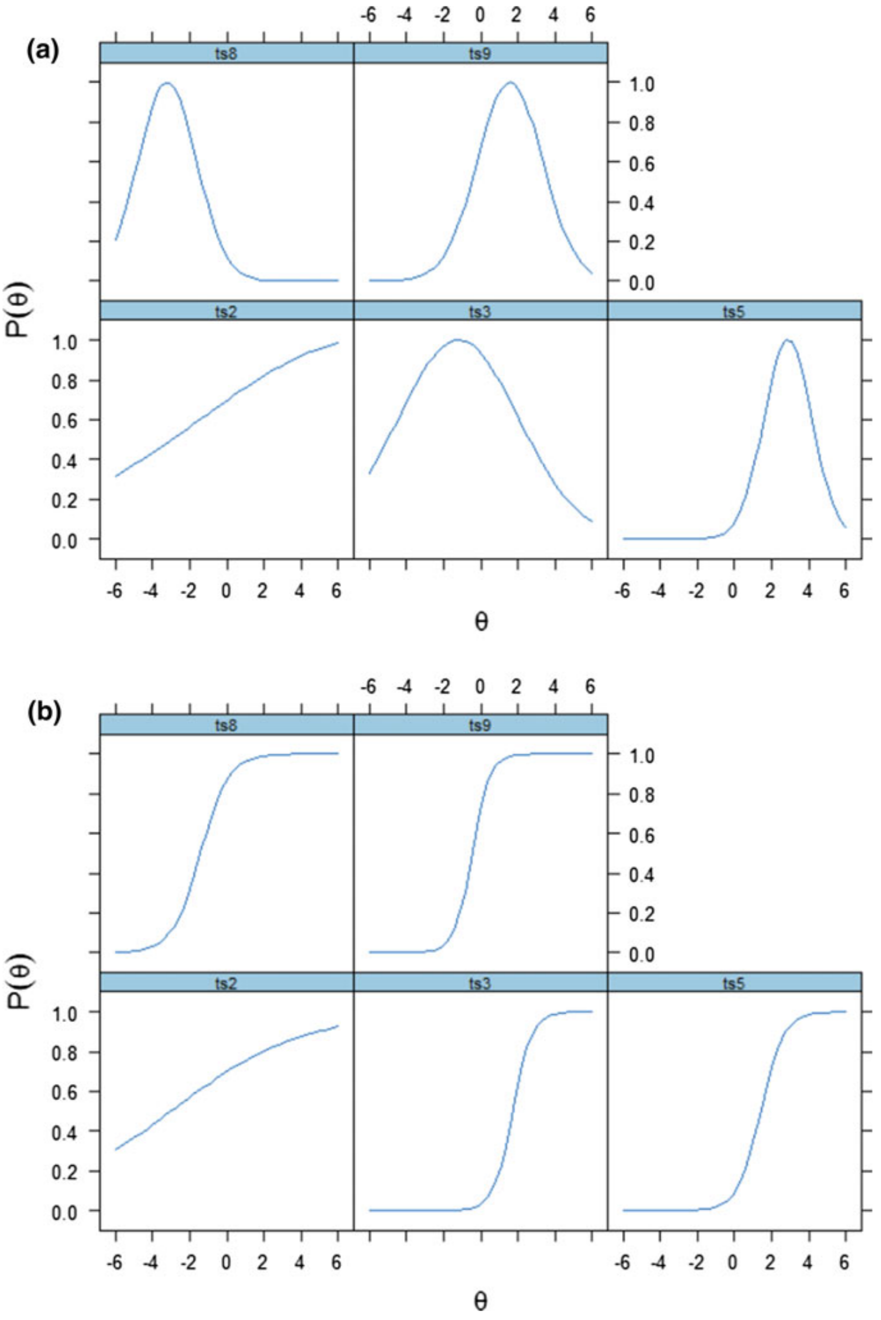


Fig. 1 Item characteristic curves for Items 2, 3, 5, 8, and 9 **a** under the ideal-point model and **b** under the two-parameter IRT model

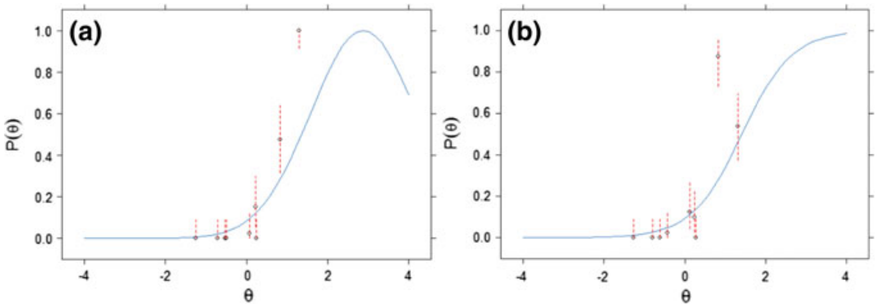


Fig. 2 Empirical observed versus expected plot for “Science is one source of knowledge among many” across **a** the ideal-point model and **b** the two-parameter IRT model

identify items that should be discarded. Indeed, despite using less items, the theta estimates from both IRT models demonstrated very similar, and at one time stronger correlations with other items compared to the classical test theory and Thurstone scaling approaches. This provides an advantage to IRT in being able to identify a fewer set of items for use, which can have positive implications for reducing survey length and subsequent respondent fatigue.

References

Borman, W. C., Penner, L. A., Allen, T. D., & Motowidlo, S. J. (2001). Personality predictors of citizenship performance. *International Journal of Selection and Assessment, 9*(1–2), 52–69.

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment, 19*(1), 88–106.

Chernyshenko, O. S., Stark, S., Prewett, M. S., Gray, A. A., Stilson, F. R., & Tuttle, M. D. (2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: Empirical comparisons with other formats. *Human Performance, 22*(2), 105–127.

Coombs, C. H. (1964). *A theory of data*. New York: Wiley.

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2009). Test theory and personality measurement. *Oxford Handbook of Personality Assessment*, 59–80.

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology, 3*(4), 465–476.

Funk, C., & Rainie, L. (2015). *Americans, politics, and science issues*. Retrieved from: http://www.pewinternet.org/2015/07/01/americans-politics-and-science-issues/.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 5–53.

Maydeu-Olivares, A., Hernández, A., & McDonald, R. P. (2006). A multidimensional ideal point item response theory model for binary data. *Multivariate Behavioral Research, 41*(4), 445–472.

McCright, A. M. (2010). The effects of gender on climate change knowledge and concern in the american public. *Population and Environment, 32*(1), 66–87. https://doi.org/10.1007/s11111-010-0113-1.

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology, 77*(4), 531–551.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X^2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*(4), 289–298.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*(1), 3–32.

Roberts, J. S., & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement, 20*(3), 231–255.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91*(1), 25–39.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*(4), 529–554.

Thurstone, L. L. (1929). Theory of attitude measurement. *Psychological Review, 36*(3), 222–241.

Thurstone, T. G. (1932). The difficulty of a test and its diagnostic value. *Journal of Educational Psychology, 23*(5), 335.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Rumor Scale Development

Joshua Chiroma Gandi

Abstract Rumor refers to unsubstantiated story or information being circulated. Although the more the integrity of the source implies the more the reliability of rumor, not all that seems reliable would be adjudged as valid. There has been cogent need for rumor validity assessment, but dearth of construct-relevant scale hampers empirical data collection. Considering that psychological scales are indispensable for assessment, the present study developed a suitable and psychometrically sound scale, using cross-sectional design and 570 randomly sampled participants. The psychometric properties are based on reliability and validity. Reliability ($\acute{\alpha} = 0.78$) was determined by item-total statistics while validity was based on content validity indexes, principal component analysis and the compatibility of factor model to the data. Seven extracted factors accounted for 92% of the total scale variance. Rumor intensity score ($R = 80$) corroborated the scale suitability. However, although the newly developed 50-item Rumor Scale is suitable for adaptation among different populations at various settings, there is need for confirmatory factor analysis (CFA) which was not implemented in the initial scale development study. Further validations, suggested to include cross-cultural and trans-national adaptations using CFA and other competing analysis models, can help to establish sufficient norms.

Keywords Validity · Scale development · Rumor scale · Gandi psychometric model · Construct-relevant scale

J. C. Gandi (✉)
Department of General and Applied Psychology, University of Jos, Jos, Plateau State, Nigeria
e-mail: gandij@unijos.edu.ng

The Psychometric Laboratories, P. O. Box 1550, Jos 930150, Plateau State, Nigeria

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_37

1 Introduction

1.1 Background to the Study

We often hear people say “it is a rumor”, “stop spreading rumors” etc. Conceptually, rumor refers to any unverified information being circulated among people to make sense of an unclear situation or to manage any threat or potential threats (Matsumoto, 2009). Rumor, according to Allport & Postman (1947), Matsumoto (2009) and Rosnow (1991), aims at making sense of an ambiguous situation or even managing physical and psychological threats. The different types of rumors are grouped based on subject matter as well as the rumor public and their object of collective concern (Matsumoto, 2009). The definition of choice, as used in the present study, refers to rumor as an unsubstantiated story or information being circulated consciously and or unconsciously based on significant intensity (Gandi, Ibadin, Musa, Gana, & Saleh, 2018).

1.2 Statement of Problem and Purpose of the Study

Rumor is capable of affecting individuals or even societal values positively or negatively (Willis, 2015). Although the more the integrity of rumor source the more the perception of its reliability, not everything that is reliable would be adjudged as valid (Cabera-Nguyen, 2010; Rao & Sinharay, 2007). There is cogent need for rumor validity empirical assessment but dearth of ‘construct-relevant’ and suitable scale hampers it. Achieving such a need requires a more optimal model, not a nomenclatural misnomer, to effectively drive the development of a suitable ‘construct-relevant’ scale. The study, which developed suitable scale for rumor validity assessment, adopted Gandi Psychometric Model (Gandi, 2018) on which the scale’s conceptual framework was designed and operationalized.

1.3 Conceptual Framework Based on Gandi Psychometric Model

The conceptual framework, in this case (Fig. 1), answered the “how and why” questions. It sets out how main stages through which the scale development process moves, i.e. from left to right. It also reflects systematic sequence of the process, from test conceptualization to scale validation, including “why” particular stage(s) or variable(s) precedes or succeeds the other (Gandi, 2018).

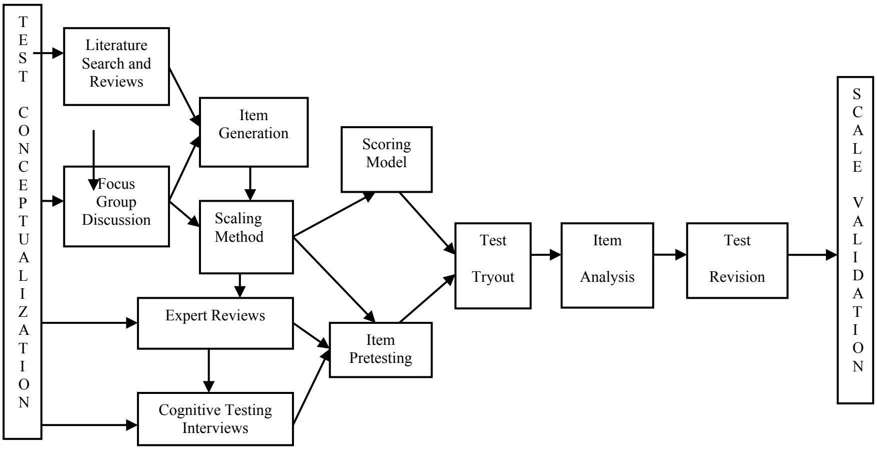


Fig. 1 Scale development framework based on Gandi psychometric model

2 Methods

2.1 Research Design and Study Setting

A cross-sectional design, considered suitable for scale development (Rivers, 2018), was adopted. It captures specific point in time, proves/disproves assumptions, applies to various types of research, and can create new theories for in-depth research. The study was conducted at three settings which include a higher institution, a public organization, and a community neighborhood setting within Jos, the capital city of Plateau State, Nigeria. Research confederates and conditions within the study setting helped to not only gauge but to also adequately prevent and avoid any perceived social desirability or other unwanted influence(s) that could amount to raping the psychometric quality of the scale in any way.

2.2 Target Population and Sample Participants

The study target population consists of members of staff at the selected higher institution and public organization as well as adult members of the selected community. They included males and females with diverse religious affiliations, ethnic backgrounds, and professional/occupational statuses. The sample size was determined by applying the principle of person-to-item ratio (i.e. subject-variable ratio). Considering the lowest and highest ratios ever used (4:1 and 36:1) respectively and the requirement of using not less than 10:1 for scale pilot-testing (Coaley, 2010; Gregory, 2000; Rosnow, 1991), the study used person-to-item ratio of 15:1. Since the scale has 50 items, the determined pilot sample size was 750 (i.e. 15 × 50). However, only 570, including 52% male (n = 296) and 48% female (n = 274), actually participated in the pilot study. The participants were selected by a stratified random sampling method. This method helped to ensure adequate representativeness by using stratification factors which include gender, religion, marital status, employment types, and occupation in the sampling process.

2.3 Materials and Procedure

Materials. The required materials for implementing a comprehensive process of scale development include relevant instruments and conditions in each case. Materials used as relevant instruments include informed consent forms, interview schedule forms, demographic data forms, focus group discussions checklist, expert reviews rating rubrics, and cognitive testing feedback sheets. Others were video camera, writing materials, SPSS software, and the processing and analysis system (computer). The materials used as conditions include the basic necessary and sufficient conditions as well as a great deal of miscellaneous (Mackie, 1965) *inus* conditions, i.e. "**i**nsufficient but **n**on-redundant part of **u**nnecessary but **s**ufficient conditions (***inus*** conditions)", based on need.

Procedure. Scale development process considered the purpose and type of scale as well as suitable psychometric perspectives (Colton & Covert, 2007; Cohen, Swerdlik & Sturman, 2013; DeVellis, 2012). It was based on nine stages of the Gandi psychometric model which include test conceptualization, item generation, scaling methods, item pretesting, scoring models, test tryout, item analysis, test revisions, and scale validation (Gandi, 2018).

Test Conceptualization. Conceiving an idea of the focal construct and what differentiates the proposed scale from others was cogent. Gregory (2000) and Cripps (2017) suggest that initial thoughts (or self-talk), such as "there ought to be a test designed to measure …", which led to creative and empirical conceptualization of the Rumor Scale for measuring rumor validity. Confronted questions, as Cohen et al. (2013) observed, include "what is the objective, content, and format of the Rumor Scale?" Thinking-aloud-questions of this kind paved the way towards reviewing rumor literature which lends credence to the scale objective, content and format. Rumor was therefore conceptualized as an unsubstantiated story or information being circulated consciously and or unconsciously based on significant intensity (Gandi et al., 2018). Equation 1, based on the formula by Allport and Postman (1947), determines rumor intensity as a mathematical function:

$$\mathrm{R} \sim \mathrm{i} \times \mathrm{a} \tag{1}$$

where "R" is the rumor intensity which includes the reach and duration of the rumor, "i" is the importance of the information in the rumor (to hearers or readers), and "a" is the ambiguity (or uncertainty) associated with the rumor and pictures the persistence of particular pieces of gossip. Rumor intensity depends on the multiplicative relationship between its importance and ambiguity.

Item Generation. Initial 38 items were deductively derived by conducting literature search and reviews (Fig. 1) as relates to the focal construct (which is “rumor”). This process also provided further useful insights on the type of statements readily comprehensible across the diverse study populations. Additional 36 items were also inductively devised by conducting focus group discussions (FGD) with a cross-section of the target population members referred to as FGD panelists (Colton & Covert, 2007; Gandi, 2018). The 74 items were appropriately worded in line with factors that constitute the term “rumor” and systematically formatted, by logical question ordering, for suitability. Item selections were done based on the item content validity index derivable from content validity ratio (CVR) of the subsequent retained items in each case. The CVR for each item are calculated using the computation formula by Lawshe (1975) as follows:

$$CVR = \frac{ne - \frac{N}{2}}{\frac{N}{2}} \tag{2}$$

where: ne is the number of FGD panelists who endorsed “essential”, N is the total number of entire FGD panelists who participated, and 2 is a constant. Furthermore, content validity index (CVI) is computed by finding the statistical mean of CVR values of the entire scale items as follows:

$$\text{CVI} = \frac{sum\, of\, CVR}{frequency} \tag{3}$$

Scaling Methods. A suitable scaling method is a necessity for newly developed scale (Rao & Sinharay, 2007). The optimally suitable method fall within the category of selective response multiple choice scaling, typically the 5-point Likert scale type (Likert, 1932; Worthington & Whittaker, 2006). Response options for the scale include ‘strongly disagree, disagree, not sure, agree and strongly agree. The advantages of and rationale for choosing Likert scale include ease of scoring, high compatibility for assessment, convenience to participants, and a robust analysis that would lead to appropriate interpretations.

Item Pretesting. 74 items earlier generated items were rigorously pretested by subjecting them to expert reviews and cognitive testing interviews (Willis, 2015). The reviewers include subject matter expert (1), content expert (1), methodology expert (1), and target population experts (2). As scholars including Rao and Sinharay (2007) and Ojedokun (2016) emphasized, “experts were asked to identify items that they particularly view as important by indicating whether such items are essential, good, or poor as relates to the focal construct”. The five expert reviewers rated each item, considering the relevance and respective match (suitability) to the construct, and their reviews reduced the 74 items to 53 items. The 53 items were further subjected to cognitive testing interviews (CTI) which used 19 participants (M = male and F

= female). They include representative sample of four academic staff (M = 2, F = 2), seven senior administrative staff (M = 4, F = 3), two clinical staff (M = 1, F = 1), five field staff (M = 3, F = 2), and a social worker (F = 1). After the survey completion, participants responded to a 7-item "cognitive testing follow-up sheet", in their own words. Three items were dropped while 50 were retained and considered for the next but one stage (i.e. pilot study or test tryout).

Scoring Model. Suitability and ease of scoring are unavoidable features that strengthen a good rating scale. Hence, the scoring model adopted for this study was a cumulative scoring, being the very aspect which demonstrated that "the higher the score on the test the higher the test-taker is on the ability/trait/other characteristic. Scoring models must agree with corresponding methods of scaling (Mackie, 1965). That was why the "considered advantages" of and rationale for choosing Likert scaling method included ease of scoring, high compatibility for assessment, convenience to participants, and a robust analysis that would lead to appropriate interpretations.

Test Tryout. The scale was tried out (Fig. 1) by conducting a pilot study using 570 persons, including 52% male (n = 296) and 48% female (n = 274), as participants. Consent forms in which the participants demonstrated voluntary willingness to participate by appending signatures were provided. The form(s) explain the study need and the participants' expectation while issues of their confidentialities and privacies were guaranteed. The demographic form, which consisted eight items, was administered alongside the newly developed scale to gather descriptive information on gender, age, marital status, religion, highest qualification, current employers, employment type, and department/unit. The Rumor Scale items appear self-validating as they helped to control any tendency for responses to reflect what could be presumed as being a desired effect (i.e. social desirability bias).

Item Analysis. The study maintained cogent culture of ascertaining item reliability index, difficulty index, discrimination index and validity index which ensured soundness of the scale. Item difficulty index and discrimination index were qualitatively determined by employing the stated rigorous processes of item pretesting. The statistical analysis, i.e. quantitative method, was used for reliability index and validity index of the retained items. Overall, the analyses conducted include content validity index (CVI), item-total statistics, exploratory factor analysis (EFA), and principal component analysis (PCA).

Test Revisions. Following the pilot study item analyses, the resulting outcomes were reviewed and classified into two categories. Those items (n = 50) found suitable and significant (labelled as category 1) were retained while those neither sound nor significant (labelled as category 2) were deleted (discarded).

3 Results

The results of examining the content validity, based on Eq. 3 formula, showed CVIs of 0.66, 0.69 and 0.72 (i.e. CVIs > 60) for item generation, expert review item pretesting and cognitive interviews item pretesting. The CVI of 0.66 resulted from the sample size of focus group discussion (FGD) panelists ($n = 18$) which represent frequency score in Eq. 3. The FGD (18) and the sum of CVR (12) therefore determined CVI to 0.66. Item pretesting, using expert reviewers (5) and the sum of their endorsements CVR (3.45) determined the expert reviews content validity, CVI to 0.69. The 53 items with content validity of 0.69 (out of the 74 items) were retained while 21 items with content validity below the adopted cutoff point of 0.60 were dropped. Again, item pretesting with cognitive testing participants ($n = 19$) and the sum of their item endorsements CVR (26.4) determined the cognitive testing content validity, CVI to 0.72. Based on the cognitive testing content validity (0.72), three of the items (CVI < 0.60) were dropped while 50 items (CVI > 0.60) were retained. Overall, content validity results shows 0.66, 0.69 and 0.72 based on item generation, expert review pretesting and cognitive testing interview analyses.

Results of rumor intensity (R) computation of the data generated with the newly developed scale, based on Eq. 1 formula, shows $R = 0.8$ (i.e. decimal of $R = 80$). This is because the rumor importance (i) score was 10 while its ambiguity (a) score was 8. Just as Eq. 1 noted that $R \sim i \times a$, which means rumor intensity is the product of its importance to the hearer (or reader) and its corresponding ambiguity, having the importance (i) as 10 and the ambiguity (a) as 8 implies that the intensity (R) is 80. It must therefore be noted that rumor intensity was computed according to percentage model and the product ($R = 80$) is converted to a decimal (i.e. $R = 80 = 0.8$).

The 50 retained scale items, resulting from significant CVIs and construct-relevant rumor intensity (R), explored had significant minimum mean (2.50) and maximum mean (4.67) as well as the lowest standard deviation and highest standard deviation (0.98 and 1.81 respectively). None of the retained items had absolute values of more than 0.30 for skewness and or kurtosis.

Internal consistency reliability based on Cronbach's alpha was $\acute{\alpha} = 0.78$ and the range of item correlation coefficients, based on Pearson, $r = 0.45{-}0.71$ ($p < 0.01$). The results, based on item-total statistics, indicated that all the items have significant reliability. The analysis, as presented in Table 1, helped to check for items that were inconsistent with the average behavior of other items on the scale. The reliability (if item deleted) was based on scale mean, scale variance, corrected item-total correlation, and Cronbach's alpha (Table 1).

Scale mean refers to if item 1 (for instance) was removed and the remaining items on the scale summed up for all 570 cases then 23.10 would be the mean of the summated items. Scale variance shows that if item 1 (for instance) was removed and the remaining items on the scale summed up for all 570 cases, then 23.02 would be the variance of the summated items. In case of the corrected item-total correlation, the correlation between (for instance) item 1and the summated score was 0.71, as presented in the third column of Table 1. Cronbach's alpha refers to if item 1 (for instance) was removed the Cronbach's alpha reliability coefficient would be 0.7744. By investigating total correlations (see Table 1), only few items have been deleted for failure to satisfy the requirements to be retained due to their low correlations while 50 items that satisfied such requirements are retained. This further improved the scale reliability which corroborated the internal consistency reliability (0.78).

Table 1 Item-total statistics (showing "if item deleted")

		Scale mean	Scale variance	Corrected item-total correlation	Cronbach alpha (ά)
1	Uncertainty about a situation leads to rumor transmission	23.10	23.02	0.71	.67
2	Rumor keeps one alert and helps to avoid danger or insecurity	23.68	21.13	0.60	.71
3	Lack of evidence is a key defining feature of rumor	23.56	22.52	0.54	.68
4	A rumor may be like "I heard that the real reason those in charge were forced to step down include *certain controversial comments* they made"	23.13	23.10	0.61	.69
5	Rumor refers to unverified information being circulated	23.51	20.49	0.48	.77
6	Information from a credible source is more believable than those spread by non-credible sources	23.60	22.12	0.45	.73
7	Reducing uncertainty or anxiety helps in managing/preventing rumors	23.06	21.45	0.66	.76
8	Information from an uncertain source destabilizes my decision making	23.22	22.04	0.59	.70

(continued)

Table 1 (continued)

		Scale mean	Scale variance	Corrected item-total correlation	Cronbach alpha (ά)
9	Rumor affects performance positively	23.17	20.23	0.68	.63
10	Rumor is an attempt to make sense of unclear situation or to manage threats	23.10	23.02	0.71	.71
11	Rumor is different from gossip in every sense	23.68	21.13	0.60	.69
12	It is okay to share any information even without evidence supporting them	23.56	22.52	0.54	.67
13	Memory limits can reduce rumor transmission	23.13	23.10	0.61	.68
14	I will not share information that seems not to have any evidence	23.51	20.49	0.48	.70
15	Uncertain events are believed to be sources of anxiety	23.60	22.12	0.45	.74
16	Rumors in agreement with my current attitudes are more believable than those that disagree	23.06	21.45	0.66	.66
17	Reducing belief in a rumor will contribute to managing it	23.22	22.04	0.59	.71

(continued)

Table 1 (continued)

		Scale mean	Scale variance	Corrected item-total correlation	Cronbach alpha (ά)
18	Rumor may read "it is believe that the main opposition party is behind the ongoing uncompromised industrial disharmony against government"	23.13	23.10	0.61	.72
19	Rumor affects performance negatively	23.51	20.49	0.48	.68
20	Anxiety promotes rumor among individuals	23.60	22.12	0.45	.68
21	Any news that turns out to become a lie, after verification, is a rumor	23.06	21.45	0.66	.67
22	Information from unreliable source has no guarantee in life	23.22	22.04	0.59	.72
23	Uncertain information constitutes more uneasiness or significant stress	23.17	20.23	0.68	.62
24	A rumor may read "have you heard? … that the businessman in town is rumored to be the financier of terrorists"	23.10	23.02	0.71	.67

(continued)

Table 1 (continued)

		Scale mean	Scale variance	Corrected item-total correlation	Cronbach alpha (ά)
25	Ambitious anticipation is a fertile ground for rumor to thrive	23.13	23.10	0.61	.64
26	A rumor being circulated repeatedly is more believable to me than those heard just once	23.51	20.49	0.48	.67
27	Rumor spreads like news if it is still new or current	23.60	22.12	0.45	.68
28	Rumor is not like news if it is never verified	23.06	21.45	0.66	.72
29	Rumor can damage an individual's reputation	23.22	22.04	0.59	.71
30	Rumor may endear the speaker to the listener	23.17	20.23	0.68	.69
31	Rumor becomes a fact whenever it is found to be true	23.10	23.02	0.71	.69
32	I feel relieved when an ambiguous situation suddenly becomes clear	23.68	21.13	0.60	.66
33	There is something exciting about being kept in suspense	23.56	22.52	0.54	.61

(continued)

Table 1 (continued)

		Scale mean	Scale variance	Corrected item-total correlation	Cronbach alpha (ά)
34	Attempts to fact finding attracts rumor mongering	23.13	23.10	0.61	.67
35	A rumor refers to any well-orchestrated propaganda being dished out	23.51	20.49	0.48	.66
36	Lack of having all the information I need frustrates me	23.60	22.12	0.45	.69
37	Rumor can damage a company's reputation	23.06	21.45	0.66	.70
38	Building trust will contribute to preventing and managing rumors	23.22	22.04	0.59	.75
39	Rumor is an evaluative chat about individuals which may or may not be verified	23.13	23.10	0.61	.65
40	Rumor keeps people on their toes for effective prevention or management of threats	23.51	20.49	0.48	.64
41	A rumor is an idle or slanderous talk about an individual who is not present	23.60	22.12	0.45	.67

(continued)

Table 1 (continued)

		Scale mean	Scale variance	Corrected item-total correlation	Cronbach alpha (ά)
42	Belief system determines the strength of rumor transmission	23.06	21.45	0.66	.77
43	Social status contributes to rumor transmission	23.22	22.04	0.59	.75
44	Rumor can foster hatred among persons or groups	23.17	20.23	0.68	.71
45	A story that brings uncertainty easily disorganizes me	23.10	23.02	0.71	.67
46	Falsification of certain rumors can reduce my belief in similar brand of rumors	23.13	23.10	0.61	.64
47	I get worried if any sensitive information I was told remains uncertain	23.51	20.49	0.48	.67
48	Rumor may read “be informed that mindless terrorists are planning attacks in town next week”	23.60	22.12	0.45	.68

(continued)

Table 1 (continued)

		Scale mean	Scale variance	Corrected item-total correlation	Cronbach alpha (ά)
49	Enhancing appropriate communication can prevent rumors	23.06	21.45	0.66	.72
50	Rumor may be circulated as "I heard that 45 members of staff will be laid off from this organization"	23.22	22.04	0.59	.71

Exploratory analysis and principal component analysis extracted seven factors with respective item membership of 10, 8, 7, 10, 6, 5 and 4 which accounted for 92% of the total scale variance (Table 2). These factors are factor 1 (rumor diagnostics), factor 2 (rumor transmission), factor 3 (rumor impact/effect), factor 4 (response attitude), factor 5 (rumor belief role), factor 6 (rumor type), and factor 7 (rumor management). The Rumor Diagnostics Subscale consists of 10 items (3, 5, 10, 11, 21, 28, 31, 35, 39 and 44), Rumor Transmission Subscale has 8 items (1, 13, 20, 25, 27, 34, 42 and 43), Rumor Impact Subscale has 7 items (2, 9, 19, 29, 30, 37 and 40), Rumor Response Attitude Subscale consists 10 items (8, 12, 14, 22, 23, 32, 33, 36, 45 and 47), Rumor Belief Subscale has 6 (6, 15, 16, 26, 41 and 46), Rumor Type Subscale has 5 items (4, 18, 24, 48 and 50), and the Rumor Management Subscale consists of 4 items (7, 17, 38 and 49).

4 Discussion

Psychological measurements are largely based on the nearest approximation of the phenomenon being measured while psychometric scales are indispensable tools needed for conducting such measurements. Scale development researchers, including Morgado, Meireles, Neves, Amaral and Ferreira (2017), however decried limitations resulting from previous scale developments which must be overcome to ensure more optimal scale. The present study essentially developed a suitable scale, referred to as the "Rumor Scale", for an empirical rumor validity assessment based on robust construct-relevant methodology (Gandi et al., 2018). Required optimality was achieved by systematic and rigorous implementation of the empirically designed Gandi Psychometric Model's 9-stage process (Gandi, 2018). This made the Rumor

Table 2 Factor loadings

		Factors						
		1	2	3	4	5	6	7
1	Uncertainty about a situation leads to rumor transmission		.61					
2	Rumor keeps one alert and helps to avoid danger or insecurity			.71				
3	Lack of evidence is a key defining feature of rumor	.60						
4	A rumor may be like "I heard that the real reason those in charge were forced to step down include *certain controversial comments* they made"						.60	
5	Rumor refers to unverified information being circulated	.88						
6	Information from a credible source is more believable than those spread by non-credible sources					.65		
7	Reducing uncertainty or anxiety helps in managing/preventing rumors							.97
8	Information from an uncertain source destabilizes my decision making				.96			
9	Rumor affects performance positively			.57				
10	Rumor is an attempt to make sense of unclear situation or to manage threats	.76						
11	Rumor is different from gossip in every sense	.69						
12	It is okay to share any information even without evidence supporting them				.77			
13	Memory limits can reduce rumor transmission		.78					
14	I will not share information that seems not to have any evidence				.71			
15	Uncertain events are believed to be sources of anxiety					.72		
16	Rumors in agreement with my current attitudes are more believable than those that disagree					.76		
17	Reducing belief in a rumor will contribute to managing it							.74
18	Rumor may read "it is believe that the main opposition party is behind the ongoing uncompromised industrial disharmony against government"						.71	
19	Rumor affects performance negatively			.99				

(continued)

Table 2 (continued)

		Factors						
		1	2	3	4	5	6	7
20	Anxiety promotes rumor among individuals		.97					
21	Any news that turns out to become a lie, after verification, is a rumor	.75						
22	Information from unreliable source has no guarantee in life				.57			
23	Uncertain information constitutes more uneasiness or significant stress				.76			
24	A rumor may read “have you heard? … that the businessman in town is rumored to be the financier of terrorists”						.88	
25	Ambitious anticipation is a fertile ground for rumor to thrive		.85					
26	A rumor being circulated repeatedly is more believable to me than those heard just once					.56		
27	Rumor spreads like news if it is still new or current		.95					
28	Rumor is not like news if it is never verified	.76						
29	Rumor can damage an individual’s reputation			.57				
30	Rumor may endear the speaker to the listener			.95				
31	Rumor becomes a fact whenever it is found to be true	.60						
32	I feel relieved when an ambiguous situation suddenly becomes clear				.91			
33	There is something exciting about being kept in suspense				.73			
34	Attempts to fact finding attracts rumor mongering		.85					
35	A rumor refers to any well-orchestrated propaganda being dished out	.71						
36	Lack of having all the information I need frustrates me				.94			
37	Rumor can damage a company’s reputation			.64		.63		
38	Building trust will contribute to preventing and managing rumors							.94
39	Rumor is an evaluative chat about individuals which may or may not be verified	.83						
40	Rumor keeps people on their toes for effective prevention or management of threats			.70				

(continued)

Table 2 (continued)

		Factors						
		1	2	3	4	5	6	7
41	A rumor is an idle or slanderous talk about an individual who is not present					.78		
42	Belief system determines the strength of rumor transmission		.90					
43	Social status contributes to rumor transmission		.82					
44	Rumor can foster hatred among persons or groups	.79						
45	A story that brings uncertainty easily disorganizes me				.94			
46	Falsification of certain rumors can reduce my belief in similar brand of rumors					.95		
47	I get worried if any sensitive information I was told remains uncertain				.95			
48	Rumor may read "be informed that mindless terrorists are planning attacks in town next week"						.99	
49	Enhancing appropriate communication can prevent rumors							.85
50.	Rumor may be circulated as "I heard that 45 members of staff will be laid off from this organization"						.89	

Scale more construct-relevant and psychometrically sound than other scales not subjected to such psychometric rigors.

Structurally, the study conceptual framework which shows empirical systematism led to answering the "how" and "why" questions in the process of devising/deriving needed items is of essence. The numerous materials adopted and used have been helpful in effective generation and efficient tracking of data for evidential documentation accuracy and analysis precision. Mackie (1974), as corroborated by Gandi et al. (2018), believe that both necessary and *inus* conditions used have effectively complemented those material instruments. Working with strictly representative sample of participants, as earlier observed (Morgado et al., 2017; Tabachnick & Fidell, 2007), has been rewarding in every sense as it contributes to ensuring that the newly developed Rumor Scale is optimally construct-relevant (Gandi et al., 2018). The implemented person-to-item ratio (i.e. subject-variable ratio) model at high significance level have added value to the study sample optimality.

Using PCA does not imply being unmindful of the suggestion, as extended by Cripps (2017), to avoid using PCA as a precursor to confirmatory factor analysis (CFA). The uniqueness of using PCA at the scale development stage, in this case, is that the use of CFA was not yet required at this level of the study (Gandi, 2018). Reliability was established based on findings from item-total statistics which checked items that were inconsistent with the average behavior of other items. Critical analysis and inspection of the scale mean, scale variance, and Cronbach's alpha if item deleted provided more insight which helped to avoid construct-irrelevant items by deleting them.

Pretesting based on cognitive testing interviews, which helped in revising and modifying certain items, has corroborated the improved suitability of the scale items among target population. For Chi-Square tests for model goodness of fits, both parallel test and strict parallel test results presented the Rumor Scale as a significant instrument that it is (Zhu & Liu, 2017). The assumption that the r population means are not equal but all the r factor loadings and population unique factor variances are equal was supported by the parallel test, while the assumption that all the r population means, factor loadings, and unique factor variances was supported by strict parallel test. The significant rumor intensity level elicited by the Rumor Scale, which supported reliability and validity, is a complementary empirical evidence of adequate suitability for assessing rumor validity. This supports Gandi et al. (2018) who opine that even as rumor predicts good news, it equally triggers distress. However, any distress resulting from conflicts and or trauma is avoidable if knowledge, skills, other resources and support are appropriately deployed (Gandi & Wai, 2010). The Rumor Scale elicits data that leads to such useful information.

Empirically, as Coaley (2010) emphasizes, the three major sources of validity evidence are based on the content issues (content-related validity), the relationships with measures of other variables (criterion-related validity), and the internal structure of the selection procedure (construct-related validity). However, validity of the present scale was based more on content validity and construct validity at its development phase. The main limitation was lack of using confirmatory factor analysis in the present study (CFA). The scale will be further validated to establish norms by implementing the CFA and other competing analysis models. Further validations, including cross-cultural and trans-national adaptations, which can sufficiently establish diverse norm(s) are suggested.

References

Allport, G. W., & Postman, L. (1947). *The psychology of rumor*. New York: Holt.

Cabrera-Nguyen, P.(2010). Author guidelines for reporting scale development and validation results. *Journal of the Society for Social Work and Research, 1*(2), 99–103.

Coaley, K. (2010). *An introduction to psychological assessment and psychometrics*. London, England: Sage.

Cohen, R. J., Swerdlik, M. E., & Sturman, E. D. (2013). *Psychological testing and assessment: An introduction to tests and measurement* (8th ed.). New York: McGraw-Hill.

Colton, D., & Covert, R. W. (2007). *Designing and constructing instruments for social research and evaluation*. London: Wiley.

Cripps, B. (2017). *Psychometric testing: Critical perspectives*. USA: Wiley Blackwell.

DeVellis, R. F. (2012). *Scale development theory and applications: Applied social research methods series*. Washington DC: Sage.

Gandi, J. C. (2018). *Development and validation of health personnel perceived quality of life scale (Unpublished doctoral thesis)*. Nigeria: University of Ibadan.

Gandi, J. C., Ibadan, J., Musa, B., Gana, J. G., & Saleh, D. A. (2018, November). *Perceived Impact of Rumor in Coexistence among Ethnically Diverse Community*. Paper session presented at the Annual Scientific Conference of Nigerian Psychological Association: Lagos Nigeria.

Gandi, J. C., & Wai, P. S. (2010). Impact of partnership in coping in mental heal recovery: An experimental study at the Federal Neuro-Psychiatric Hospital Kaduna. *International Journal of Mental Health Nursing, 19*(5), 322–330.

Gregory, R. J. (2000). *Psychological testing: History, principles, and applications*. Boston: Allyn and Bacon.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28,* 563–575.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140.

Mackie, J. L. (1965). Causes and conditions. *American Philosophical Quarterly, 12,* 245–265.

Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Oxford England: Oxford University Press.

Matsumoto, D. (2009). *The Cambridge dictionary of psychology*. United Kingdom: Cambridge University Press.

Morgado, F. F. R., Meireles, J. F. F., Neves, C. M., Amaral, A. C. S., & Ferreira, M. E. C. (2017). Scale development: ten main limitations and recommendations to improve future research practices. *Psicologia: Reflexao E Criyica, 30*(3), 1–20.

Ojedokun, O. (2016). Development and psychometric evaluation of the littering prevention behavior scale. *Ecopsychology, 8*(2), 138–152.

Rao, C. R., & Sinharay, S. (2007). Handbook of statistics on psychometrics, Vol. 26.

Rivers, J. (2018). Cross-sectional study: Definition, advantages, disadvantages & example. Retrieved from https://study.com/academy/lesson/cross-sectional-study on September 22, 2018.

Rosnow, R. L. (1991). Inside rumor: A personal journal. *American Psychologist, 46*(5), 484.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Pearson Education: Fifth Edition.

Willis, G. B. (2015). *Analysis of the cognitive interview in questionnaire design*. Oxford University Press.

Worthington, R., & Whittaker, T. (2006). Scale development research: A content analysis and recommendation for best practices. *Counselling Psychologists, 34,* 806–838.

Zhu, X., & Liu, F. (2017). Research on behavior model of rumor based on systems dynamics. *Hindawi Complexity, 2,* 1–9.

An Application of a Topic Model to Two Educational Assessments

**Hye-Jeong Choi, Minho Kwak, Seohyun Kim, Jiawei Xiong,
Allan S. Cohen and Brian A. Bottge**

Abstract A topic model is a statistical model for extracting latent clusters or themes from the text in a collection of documents. The purpose of this study was to apply a topic model to two educational assessments. In the first study, the model was applied to students' written responses to an extended response item on an English Language Arts (ELA) test. In the second study, a topic model was applied to the errors students' made on a fractions computation test. The results for the first study showed five distinct writing patterns were detected in students' writing on the ELA test. Two of the patterns were related to low scores, two patterns were associated with high scores and one pattern was unrelated to the score on the test. In the second study, five error patterns (i.e., latent topics) were detected on the pre-test and six error patterns were detected on the post-test for the fractions computation test. The results for Study 2 also yielded evidence of instructional effects on students' fractions computation ability. Following instruction, more students in the experimental instruction condition made fewer errors than students in the business-as-usual condition.

Keywords Topic models · Extended response items · Error analysis

1 Introduction

Topic models are statistical models for extracting the latent themes or topics from the texts in a collection of documents that best represents the thematic structure in the collection (Blei, 2012; Blei et al., 2003; Griffiths & Steyvers, 2004). Topic modeling has been widely used in a variety of areas. Lauderdale (2014), for example, used a topic model for studying the policy orientation of a court's opinions, and Grimmer (2010) used it for studying priorities of legislators and other political actors. Lau et al. (2012) used a topic model for analyzing thematic trends on twitter. Bisgin et al.

H.-J. Choi (✉) · M. Kwak · S. Kim · J. Xiong · A. S. Cohen
University of Georgia, Athens, USA
e-mail: hjchoi1@uga.edu

B. A. Bottge
University of Kentucky, Lexington, USA

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_38

(2011) used a topic model to study drug labeling with the objective of discovering topics that can group drugs together based on similar safety concerns or therapeutic uses, and Rhody (2012) discussed how a topic model can be applied for poems, or figurative language.

To date, however, little research has been reported on analysis of the text of answers to test questions on educational assessments. In educational assessments, constructed response items are proposed for use in assessing higher order thinking skills that are less amenable to testing via selected response type items (Brookhart, 2010). In their responses to these items, examinees construct answers using their own words as compared to selecting an answer choice. Scores for students' constructed responses to these items are based on rubrics. These provide an estimate of examinee ability but may not provide much information about students' higher order thinking skills. Using topic modeling, it may be possible for researchers to extract the latent themes reflecting these thinking skills from students' responses.

In this study, we applied a topic model in two studies to help in identifying the primary latent topics contained in the text of students' responses: In Study 1, we analyzed responses to a formative assessment in English Language Arts (ELA) for eighth grade students. In Study 2, we analyzed error patterns on a mathematics test assessing knowledge of fractions for middle school students. Below, we provide a brief description of a topic model and then illustrate how to apply this model to students' responses in educational assessments.

2 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA; Blei et al., 2003) is one of the simplest topic models. Suppose that a corpus consisted of D documents with each indexed by d. Each document can be presented as a vector of words, $w_d = (w_{d,1}, \ldots, w_{d,N})'$, where N is the number of words in document d.

In LDA, each document in a corpus is assumed to be a mixture of finite topics, and each topic is considered a probability distribution over a fixed set of words, or a multinomial distribution over words, with a vector of probabilities $\beta_k = (\beta_{k,1}, \ldots, \beta_{k,V})$ for $k = 1, \ldots, K$, where V is the number of unique words, and K is the number of topics. Each document has topic proportions, $\theta_d = (\theta_{(d,1)}, \ldots, \theta_{(d,K)})'$. Then, the per-document topic distribution (θ_d) and topic-words distribution (β_k) are assumed as

$$\theta_d \sim Dirichlet(\alpha), \qquad \beta_k \sim Dirichlet(\eta),$$

where α and η indicate the parameters of the Dirichlet distribution.

Each word ($w_{(d,n)}$) in a document has a topic assignment ($z_{(d,n)}$) that shows which topic the word is from, and the word is assumed to be drawn from the corresponding topic. Topic assignments for each document, $z_d = (z_{(d,1)}, \ldots, z_{(d,N)})'$, are determined by θ_d, meaning that $z_{(d,n)}$ follows a multinomial distribution over topics with probabilities of θ_d.

In LDA, the order of the words and the grammatical role of the words in the document do not matter (this is the "bag of words" assumption; Blei, 2012). The observed variables are the words of documents, and the latent or hidden variables are the topic structure (topic-words distribution, per-document topic distribution, and topic assignments). And, finally, the joint distribution of $\beta = \{\beta_1, \ldots, \beta_K\}$, $\theta = \{\theta_1, \ldots, \theta_D\}$, $z = \{z_1, \ldots, z_D\}$, and $w = \{w_1, \ldots, w_D\}$, given the parameters of α and η, can be written as follows (Blei et al., 2003; Blei, 2012):

$$P(\beta, \theta, z, w|\alpha, \eta) = \prod_{k=1}^{K} p(\beta_k|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha)[\prod_{n=1}^{N} p(w_{d,n}|z_{d,n}, \beta)p(z_{d,n}|\theta_d)], \quad (1)$$

where N is the number of word in each document which may be different across documents.

In fitting LDA, there are multiple ways of extracting the latent topic structure. Most common are by Gibbs sampling method (Steyvers & Griffiths, 2007) or by a variational method (Blei et al., 2003).

In the next section, we applied LDA to two sets of education assessments. In determining the number of topics, we fit the model with different number of topics, from 2- to 10-topic models, and DIC (Spiegelhalter, 1998) and interpretability were used for model selection. As each topic is characterized by its distribution over words, we used the most probable words or terms in the topic for interpreting and labeling each topic. To implement LDA, R packages topicmodels (Hornik & Grün, 2011) and stm (Roberts, Stewart, Tingley, 2018) were used.

3 Study 1: Argumentative Writing

3.1 Assessment and Data

In Study 1, the data were taken from a formative assessment of English and Language Arts (ELA) for Grade 8. The assessment consists of 6 items: three multiple choice, one short answer and one extended response items. The non-linear reliability (Green & Yang, 2009) of this assessment was 0.85. In this study, students' extended response to the extended item was analyzed. The item consisted of two passages: one passage was about environmental facts and the other was about economic facts. Students were instructed to write an argumentative essay indicating whether their congressional representative should allow the protected forest to be developed into commercial timberland.

Before fitting the LDA, words with a frequency of less than 10 and documents with less than 2 words were excluded. Also, stop words were eliminated from the documents as a preprocessing step as they carry little if any useful information about the topic. Stop words for this item included such terms as *am*, *is*, *are*, *a*, *the*, and *but*.

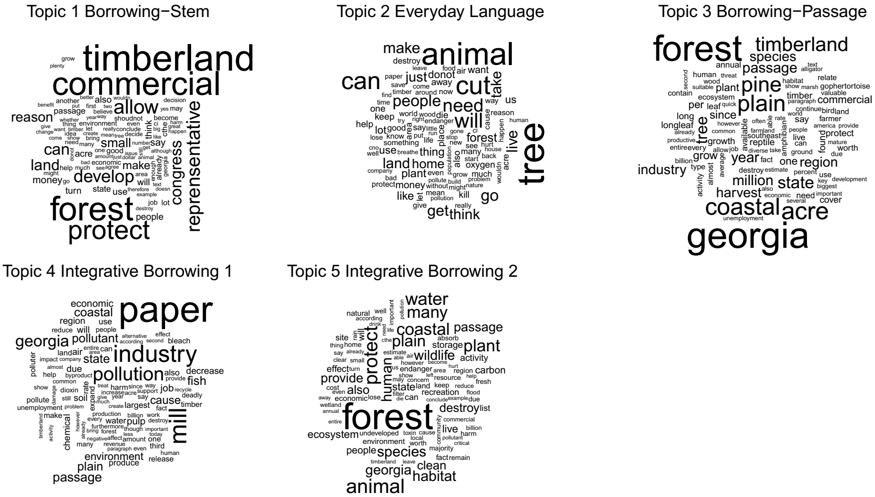
The final data set included 2230 students' responses with a total of words 257,378 in this corpus. The number of unique words was 1006 and the average answer length was 115 words (SD=75 words). After the preprocessing step, LDA models were fit with from 2 to 20 topic.

3.2 Results

DIC suggested a 5-topic model for the data. Figure 1 presents a frequency-weighted word cloud. Word clouds are commonly used to visually represent the words in a corpus. The font size of a word in Fig. 1 is proportional to its importance in the corpus.

Based on the 30 most probable words in each topic, the first and third topics were characterized as indicating simple *borrowing* of words. That is, students borrowed words from the stem for the first topic and from the passages for the second topic without using them to support their argument. The second topic was characterized as use of *everyday language*. In this topic, students used everyday language about the topic but did not necessarily provide detailed support for their arguments.

The fourth and fifth topics were characterized as *integrative borrowing*. This was actually what the item requested students to do. That is, students were specifically instructed in the prompt to use words from the stem or the passages to support their arguments. These two topics were characterized as well-integrated answers that used words to present and support students' arguments.


Topic 1 Borrowing−Stem

Topic 2 Everyday Language

Topic 3 Borrowing−Passage

Topic 4 Integrative Borrowing 1

Topic 5 Integrative Borrowing 2

Fig. 1 Word clouds for individual topics containing the 30 most frequent words in students' writing for the ELA extended response item. The labels for individual topics are on top of individual word clouds

Correlations were calculated between the individual topics and the rubric based scores: *integrative borrowing* topics were significantly and positively related to scores (r = .31 for Topic 4, r = .25 for Topic 5); however, simple *borrowing* or using *Everyday Language* were negatively related to scores. In particular Topic 2 , *Everyday Language*, was significantly and negatively related to scores (r = −.28). The following are examples of students' answers that are representative of each topic. The five most frequent words are underlined for each topic.

Topic 1 *Borrowing-Stem*: "I do not believe that the congressional representative should allow a small protected forest to become commercial timberland. I say this because in the passage it quotes "Commercial timberland has acres and acres of forest all over the state" So why would they need a small protected forest they really don't need the small protected forest because they already have enough land as it is. Since they have all of that land they shouldn't want a small forest for their commercial timberland. I believe we barely will have enough to survive off of if they keep cutting down all the trees for example using all the forest for commercial timberland. They are using enough of our forest as it is. They really don't even need our small protected forest."

Topic 2 *Everyday Language*: "They can make a small tree farm, but they have to be careful on what they are doing just in case if there are any living animals living in those trees. It is not fare for the animals to suffer without a home because we took their home away. What happens if someone comes and destroys your house, how would you feel, that's how the animals feel when you take their homes and make it into a pieces of paper. So they have to find a new home. Then they destroy it again and the process keeps on going and going and going."

Topic 3 *Borrowing-Passage*: "Sixty percent of Georgia's coastal plain is covered in forest. The forest is one of the most diverse ecosystems in America and includes forest, grassland, sandhills, marsh, swamp, and coastal habitats. Several varieties of pine and oak are the most common trees. The growth on the ground under the longleaf pine forest contains 150–300 plant species per acre, more birds than any other Georgia forest type, and 60% of the amphibian and reptile species found in the Southeast."

Topic 4 *Integrative Borrowing*: "Paper mills are the third largest polluters in the U.S., releasing pollutants into the air, water, and soil. Dioxin, a common paper bleaching chemical byproduct, is one of the most deadly pollutants. Though Georgia still feels the effect of pollution from decades past, many paper mills are working to reduce the amount of pollutants they produce today. Paper mills can use alternative bleaching chemicals, treat their wastewater, and support paper recycling to reduce the impact on the environment."

Topic 5 *Integrative Borrowing*: "No Georgia should not make a timberland commercial because according to passage B: There are some concerns about protecting the natural resources in the coastal plains region, however. Georgia's protected forest lands are worth an estimated $37 billion annually to Georgia by remaining undeveloped. The forests provide clean water, carbon storage, wildlife habitats, and recreation sites for many people. Passage A: The effects of human activity could be costly, not only to animals and plants but to humans as well. On Georgia's Protected Species List there are 111 animals and 103 plants, the majority of which live in the coastal plain ecosystem. The wetlands of the coastal plain reduce floods by absorbing rain water and provide fresh water for local communities; they are also able to absorb and filter some toxins and pollutants."

3.3 Summary for Study 1

The five topics detected from the ELA responses were characterized as being either simple borrowing of words from the prompt or passages and thus not responsive to the instructions for the item (i.e., Topics 1 and 3), simply using everyday language as opposed to use of the terms in the passages for the item (i.e., Topic 2), and appropriately using borrowing of words to support the answer (i.e., Topics 4 and 5). Thus, the results of the LDA analysis helped to amplify the information extracted from students' answers. It provided information about what kinds of responses were likely to have higher scores and what kinds of answers were likely to have lower scores. In this way, the topic modeling added to the information obtained from the grading of students' answers to help explain why students may have received lower scores, as was the case for answers characterized by Topics 1 to 3, or why use of Topics 4 and 5 was associated with higher scores.

4 Study 2: Errors on Fractions Computation

In the second study, we applied LDA to middle grades students' errors on a fractions computation test. As we show in this example, the results of the topic modeling illustrate how LDA can be used to help investigate differential effects of instruction on students' learning.

4.1 Assessment and Data

Data. Participants included 756 middle school students in Grades 6 to 8. 433 of the students had diagnosed learning disabilities, and the remaining 323 students were considered as average achieving in mathematics. Individual classes were randomly assigned to an experimental instructional condition (N = 360) or to a business-as-usual (BAU) condition (N = 396).

Instructional Conditions. The experimental condition implemented was Enhanced Anchored Instruction (EAI; Bottge et al., 2014). Teachers in the EAI condition asked probing questions and offered instructional guidance to students as they viewed videos about the mathematics they were learning. Teachers also helped students identify relevant information in the videos to help them solve the problem. This helped eliminate the need for reading, which is a skill many low-achieving math students also lack. In the BAU condition, students received instruction over the same mathematics as in the EAI condition, but teachers did not use supportive or probing questions nor enhance instruction using videos explaining and illustrating the mathematics being taught.

Fractions Computation Test. A Fractions Computation Test (FCT) designed by Bottge et al. (2014) consisted of 20 items (14-addition and 6-subtraction items). The FCT was administered for the pre-test and the post-test for investigating instruction effects on students' fractions computation ability. Math education experts identified 11 types of errors students made when incorrectly answering these items. A description of each error is given below:

Combining (*C*): Student combines numerators and combines denominators, consistently applying the same operation to numerator and denominator.
Select Denominator (*SD*): Student selects one of the denominators listed in the problem and makes no attempt to make equivalent fraction. Denominator given in the answer must be present in the problem.
No response (*NR*): Student leaves problem blank.
Add All (*AA*): Student separately adds together all the components of the fractions.
Equivalent Fraction Error (*EQ*): Student makes an error when attempting to represent an equivalent fraction.
Large-small (*LS*): Student subtracts smaller from larger fraction out of order. Or student subtracts smaller part of fraction from larger part of fraction out of order when combined with (*C*) error.
Computation Error (*CE*): Student makes an arithmetic error.
Adding Components (*AC*): Students adds the numerator and denominator of each individual fraction together and those two sums are represented in the answer.
Wrong Operation (*WO*): Student adds given a subtraction problem or subtracts given an addition problem.
Renaming (*RN*): Student makes a mistake when renaming a whole number as a mixed number; the student fails to borrow correctly from a whole number.
Other (*O*): Student makes error other than those listed above.

These errors reflect students' misunderstandings about computing with fractions. Table 1 presents frequencies of individual errors on the pre-test and the post-test. In that table, *No-Error* indicates that students correctly answered the item. The table shows that the most common errors were *Combining* (*C*) and *Selecting Denominator* (*SD*). The remaining nine other types of errors occurred less frequently.

4.2 Results

LDA was used to analyze the errors in this data set. Models with from 2 to 10 topics were fit to the error data. DIC suggested a 5-topic model for the pre-test and a 6-topic

Table 1 Frequency of errors students made on the pre-test and the post-test

Error code	(Quick description)	Pre-test	Post-test
NE	(No-error)	3288	6172
C	(Combining)	5290	3061
SD	(Select denominator)	1287	1270
NR	(No-response)	925	691
AA	(Add all)	564	195
EQ	(Equivalent fraction error)	475	725
LS	(Large-small)	440	530
CE	(Computation error)	150	358
AC	(Adding components)	139	88
WO	(Wrong operation)	80	109
RN	(Renaming)	23	53
O	(Other)	2459	1868
Total		15,120	15,120

model for the post-test to be the best fit of the candidate models considered. Figure 2 shows mixtures of errors in each topic using a stacked bar plot. Error codes with frequencies of greater than 20% are shown in the bar. The plot in the left panel is for the pre-test and the plot in the right panel is for the post-test.

The plots illustrate differences in types of errors made by students on each of the topics. Most topics on the pre-test and the post-test were relatively similar. Topic 1 consisted of mainly *Other*. Topic 2 consisted of *Add all* and *No-response*. Topic 3 was a mixture of *Select Denominator* and *No-error*. Topic 4 was mostly *Combining*

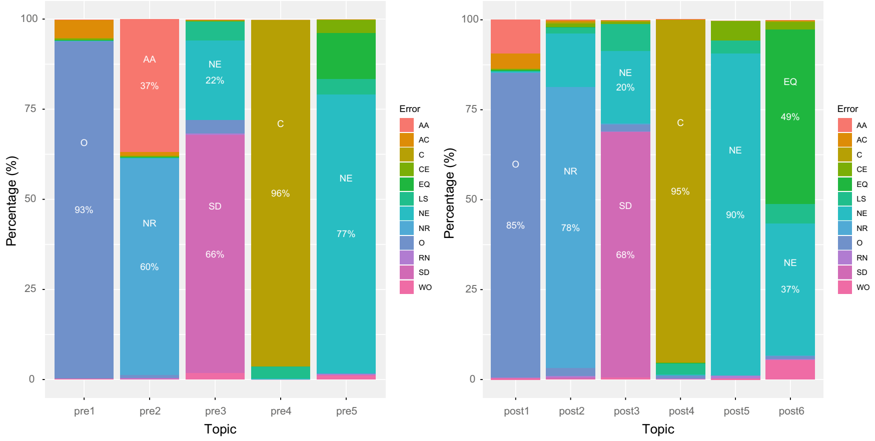


Fig. 2 Topic structure characterized by errors students made on the fractions computation test: left panel for the pre-test and right panel for the post-test

errors. Topic 5 was mainly *No-error* which is an indication of demonstrating fractions computation ability. Topic 6 only appeared on the post-test and consisted of a mixture of *Equivalent Fraction Error* and *No-error*.

For all students in this study, the averaged topic proportion for Topic 5 *No-Error* occurred in 24% of the answers on the pre-test and 35% on the post-test. This suggests that students' fractions computation ability increased following both BAU and EAI instruction.

Figure 3 presents comparisons of topic proportions averaged across students between BAU and EAI classes on the post-test. The plot in the upper panel is for students in BAU classes and the plot in the lower panel is for students in EAI classes. Distributions for Topic 5 on the post-test occurred in 47% of students' answers in the EAI group and 24% for students in the BAU group. In addition, distributions of errors in the BAU condition indicated that, even after their normal math instruction, students still made a *Combining error* about 30% of the time. In particular, students combined numerators and combined denominators incorrectly.

The increase in Topic 5, *No-Error* is indicative of the EAI effect. A beta regression analysis on Topic 5 was used to further investigate the EAI effect. Results indicated that the odds of having a higher *No-Error* topic proportion in the EAI group was 2.07 times that for the BAU group ($[exp(0.729) = 2.07(t = 9.11, p < .001, df = 753)]$. In other words, after EAI instruction, students tended to better understand the fractions computation material than students in the BAU group.

In addition to the document proportions, Table 2 presents some of the students' error patterns in the EAI condition to illustrate how students' error patterns changed from pre-test to post-test. Values in the parentheses are topic proportions for the six topics for each document indicating that how much of each topic is present in each

Fig. 3 Comparison of topic distribution between BAU and EAI classes at the post-test

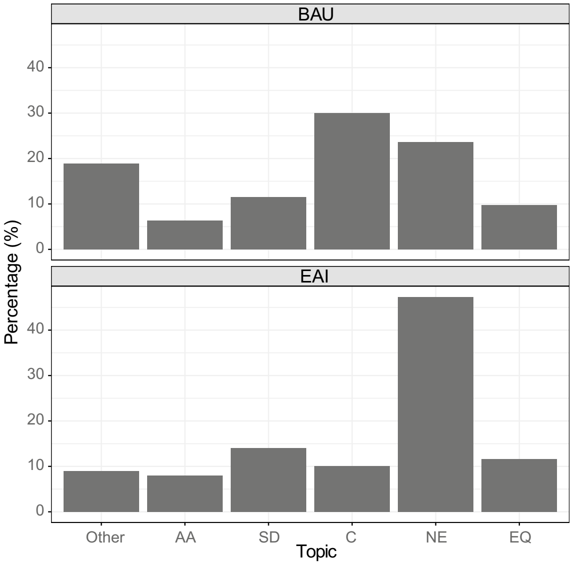


Table 2 Transitioning students' errors from pre-test to post-test

Student ID	Test (Topic proportion (%), θ_d)a	Error code
734	Pre-test (16,7,**72**,3,3, NA)	NE NE SD SD SD SD O O SD SD SD SD SD SD WO WO WO WO WO WO
	Post-test (2,2,2,2,**80**,11)	NE NE NE NE NE NE NE NE NE NE NE NE NE CE NE NE NE LS NE CE
552	Pre-test (3,3,3,**85**,7, NA)	C NE C C C C C C C C C C C C C C C LS C C
	Post-test (2,2,2,2,**90**,2)	NE NE NE NE NE NE NE NE NE NE NE NE NE CE NE NE NE RN NE LS
677	Pre-test (**55**,3,11,3,29, NA)	NE NE O O O O O O O O O O O O NE SD NE LS SD LS
	Post-test (2,2,2,2,**90**,2)	NE NE NE NE NE NE NE NE NE NE NE NE NE NE NE NE NE NE NE LS

[a]The order of topics is *Other, Add all, Select Denominator, Combining, No-error*, and *Equivalent Fraction error*. The values in bold are for the most likely topic

document. Prior to instruction, on the pre-test, student ID = 734 often made *Select Denominator* errors and *Wrong Operation* errors. After EAI instruction, the same student did not made any mistakes on 18 of the items and only made *Large-Small* and *Computation Error* errors on the remaining items. Similarly, student ID = 552 made *Combining* errors and student ID = 677 made *Other* errors on the pre-test but after instruction, neither of these kinds of errors were made on the post-test by either student.

5 Conclusions and Discussion

LDA has been shown to be a useful tool for analyzing various types of data including text, music, and images (Hu, 2009). In education assessment, the situation is different in that constructed response test items ask for more focused types of responses from students. Constructed response items, for instance, are often used to measure higher order cognitive abilities. Rubric based scoring for these items does provide useful information about the correctness of answers, but the scores alone do not provide information that helps understand student's thinking. LDA extracts latent themes in text and, as a result, can be a means of providing additional kinds of information about students' thinking as they answer constructed response items. One possibility is that this kind of information can help describe how students use words in textual answers and how this relates to better understanding the thinking they use in their answers.

In this study, we used LDA for analyzing students' writing on an ELA test and for analysis of error students made on a middle grades test of computing fractions.

LDA identified distinct topics in both ELA and fractions computation tests. Further, results for the fractions computation test showed instructional effects on students' misconceptions of fractions computation. LDA appears to have the potential to be useful for investigating students' thinking as reflected in their textual answers and patterns of errors. With respect to the patterns of errors, it was interesting to note that these tended to be reflective of the effects of the instructional conditions in Study 2.

Acknowledgements The fractions computation data used in the article were collected with the following support: the U.S. Department of Education, Institute of Education Sciences, PR Number H324A090179.

References

Bisgin, H., Liu, Z., Fang, H., Xu, X., & Tong, W. (2011, December). Mining FDA drug labels using an unsupervised learning technique-topic modeling. In *BMC bioinformatics Vol. 12*, No. 10, p. S11. BioMed Central.

Blei, D. M., Ng, A. Y., Jordan, M., & I,. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Bottge, B. A., Ma, X., Gassaway, L., Toland, M., Butler, M., & Cho, S. J. (2014). Effects of blended instructional models on math performance. *Exceptional Children, 80*, 423–437.

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. Alexandria, VA: ASCD.

Green, S. B., & Yang, Y. (2009). Reliability of Summed Item Scores Using Structural Equation Modeling: An Alternative to Coefficient Alpha. *Psychometrika*, *74*, 155–167.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228–5235.

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, *18*(1), 1–35.

Hornik, K., & Grün, B. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1–30.

Hu, D. J. (2009). *Latent dirichlet allocation for text, images, and music*. University of California, San Diego. Retrieved November, 16, 2018, from http://cseweb.ucsd.edu/~dhu/docs/research_exam09.pdf.

Lau, J. H., Collier, N., & Baldwin, T. (2012). On-line trend analysis with topic models: # twitter trends detection topic model online. *Proceedings of COLING*, *2012*, 1519–1534.

Lauderdale, B. E., & Clark, T. S. (2014). Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science*, *58*(3), 754–771.

Rhody, L. (2012). Topic modeling and figurative language. Journal of Digital Humanities, 2(1), pp. 19-35. Lauderdale, B. E., & Clark, T. S. (2014). Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science, 58*(3), pp. 754-771.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2018). stm: R Package for Structural Topic Models. http://www.structuraltopicmodel.com.

Spiegelhalter, D. J., Best, N. G., & Carlin, B. P. (1998). *Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models*. Medical Research Council Biostatistics Unit, Cambridge, UK: Technical report.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis: A road to meaning. Hillsdale* (pp. 427-448). Mahwah, NJ: Lawrence Erlbaum Associates.