

Springer Proceedings in Mathematics & Statistics

Marie Wiberg · Dylan Molenaar ·
Jorge González · Jee-Seon Kim ·
Heungsun Hwang *Editors*

Quantitative Psychology

The 86th Annual Meeting
of the Psychometric Society, Virtual,
2021

 Springer

**Springer Proceedings in Mathematics &
Statistics**

Volume 393

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including data science, operations 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.

Marie Wiberg • Dylan Molenaar • Jorge González •
Jee-Seon Kim • Heungsun Hwang
Editors

Quantitative Psychology

The 86th Annual Meeting of the
Psychometric Society, Virtual, 2021

 Springer

Editors

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

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

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

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

Heungsun Hwang
Department of Psychology
McGill University
Montreal, QC, Canada

ISSN 2194-1009

ISSN 2194-1017 (electronic)

Springer Proceedings in Mathematics & Statistics

ISBN 978-3-031-04571-4

ISBN 978-3-031-04572-1 (eBook)

<https://doi.org/10.1007/978-3-031-04572-1>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

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

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

Preface

The Covid-19 pandemic was still present in most parts of the world and thus the 86th annual meeting of the Psychometric Society was conducted virtually. This volume represents presentations given at this meeting. This is the second IMPS meeting held exclusively online and took place from July 20 to 23, 2021. There were 244 abstracts submitted (including 164 oral presentations, 53 posters, and 4 organized symposia). The virtual meeting attracted 381 participants, 67 of whom also participated in virtual short course pre-conference workshops. There was one keynote presentation, five invited presentations, five spotlight speaker presentations, one dissertation award presentation, two early career award presentations, and one career award presentation.

Since the 77th meeting in Lincoln, Nebraska, Springer publishes the proceedings volume from the annual meeting of the Psychometric Society to allow presenters at the annual meeting to spread their ideas quickly to the wider research community, while still undergoing a thorough review process. To share research through virtual presentations and written work is especially important as meeting in person is still difficult in many parts of the world in 2021. The previous nine volumes of the IMPS proceedings were received successfully, and we expect these proceedings to be successful as well.

The authors were asked to use their presentation at the meeting as the basis of their chapters. The authors also had the possibility to extend their chapters with new ideas or additional information. The result is a selection of 24 state-of-the-art chapters addressing several different aspects of psychometrics. The contents of the chapters include but are not limited to item response theory, test scores, cognitive diagnostic models, response time, psychometric models, and several applications within different fields.

Umeå, Sweden
Amsterdam, The Netherlands
Santiago, Chile
Madison, WI, USA
Montreal, QC, Canada

Marie Wiberg
Dylan Molenaar
Jorge González
Jee-Seon Kim
Heungsun Hwang

Contents

Some Adventures in Reliability Estimation	1
Anne Thissen-Roe	
Detecting Testlet Effects in Cognitive Diagnosis Models	17
Youn Seon Lim	
Comparison of Estimation Algorithms for Latent Dirichlet Allocation ...	27
Constanza Mardones-Segovia, Hye-Jeong Choi, Minju Hong, Jordan M. Wheeler, and Allan S. Cohen	
Relationship Between Students’ Test Results and Their Performance in Higher Education Using Different Test Scores	39
Marie Wiberg, Juan Li, Per-Erik Lyrén, and James O. Ramsay	
Statistical Properties of Lower Bounds and Factor Analysis Methods for Reliability Estimation	51
Julius M. Pfadt and Klaas Sijtsma	
Modeling Covarying Responses in Complex Tasks	65
Amanda Luby and Riley E. Thompson	
Exploring the Utility of Nonfunctional Distractors	83
Merve Sarac and Richard A. Feinberg	
Continuation Ratio Model for Polytomous Items Under Complex Sampling Design	95
Diego Carrasco, David Torres Irribarra, and Jorge González	
Impact of Likelihoods on Class Enumeration in Bayesian Growth Mixture Modeling	111
Xin Tong, Seohyun Kim, and Zijun Ke	
A Bi-level Individualized Adaptive Learning Recommendation System Based on Topic Modeling	121
Jiawei Xiong, Jordan M. Wheeler, Hye-Jeong Choi, and Allan S. Cohen	

Derivation of the Percentile Based Parameters for Tukey HH, HR and HQ Distributions 141
 Yevgeniy Ptukhin, Yanyan Sheng, and Todd Headrick

Predicting Human Psychometric Properties Using Computational Language Models 151
 Antonio Laverghetta Jr., Animesh Nighojkar, Jamshidbek Mirzakhlov, and John Licato

Predicting Item Characteristic Curve (ICC) Using a Softmax Classifier 171
 Dmitry I. Belov

Pooled Autoregressive Models for Categorical Data 185
 Zhenqiu (Laura) Lu and Zhiyong (Johnny) Zhang

An Investigation of Prior Specification on Parameter Recovery for Latent Dirichlet Allocation of Constructed-Response Items 203
 Jordan M. Wheeler, Jiawei Xiong, Constanza Mardones-Segovia, Hye-Jeong Choi, and Allan S. Cohen

Methods to Retrofit and Validate Q-Matrices for Cognitive Diagnostic Modeling 217
 Charles Vincent Hunter, Hongli Li, and Ren Liu

The Sum Scores and Discretization of Variables Under the Linear Normal One-Factor Model 227
 Rudy Ligtvoet

Identifying Zones of Targeted Feedback with a Hyperbolic Cosine Model 237
 Ye Yuan and George Engelhard

A Reduced Social Relations Model for Dyad-Constant Dependent Variables 249
 Terrence D. Jorgensen and K. Jean Forney

Quality Assurance in Digital-First Assessments 265
 Manqian Liao, Yigal Attali, Alina A. von Davier, and J. R. Lockwood

Effects of Restoring Missing Data on Uniform Differential Item Functioning Analysis 277
 Ya-Hui Su and Bin-Chin Lu

Two-Step Approach to Topic Modeling to Incorporate Covariate and Outcome 287
 Minju Hong, Hye-Jeong Choi, Constanza Mardones-Segovia, Yasemin Copur-Gencturk, and Allan S. Cohen

Modeling Student’s Response Time in an Attribute Balanced Cognitive Diagnostic Adaptive Testing 299
Tong Wu, Shaoyang Guo, and Hua-Hua Chang

Impact of Construct Reliability on Proposed Measures of Structural Fit When Detecting Group Differences: A Monte Carlo Examination 313
Graham G. Rifenbark

Index 329

Some Adventures in Reliability Estimation



Anne Thissen-Roe

Abstract In many measurement settings, the reliability of a measure can be straightforwardly estimated. One might have a test that is at least supposed to be unidimensional, on which everyone is supposed to respond to all the same items, and the score is a simple sum of correct answers. Or maybe one has a computerized adaptive test, producing score estimates with individual errors of measurement and corresponding individual reliability values, all based on the same posterior distribution. But sometimes, such as in industry applications of gamified assessment, one arrives at measures for which one wants to estimate the reliability—and yet, they don’t look much like traditional tests, or use IRT for scoring. Then, one is obliged to be adventurous. This collection of anecdotes from a career-matching platform features applications of a variety of techniques for estimating reliability in unusual situations, such as composite reliability, structured applications of split-half, and modeling and simulation.

Keywords Reliability · Game-based assessment

1 Background

At pymetrics, we do career matching. We’ve turned a set of classic psychology and neuroscience experiments into web and phone games. From the gameplay we get a highly multidimensional set of measures, on which we use machine learning to map users’ score patterns to job families where people like them are successful. A benefit of this approach is that we can match applicants to careers based on qualities of people from every walk of life. Most measures do not have “good” and “bad” directions, but distinguish the needs of one kind of job task from another,

A. Thissen-Roe (✉)
Pymetrics, Inc., New York, NY, USA
e-mail: anne@pymetrics.com; <https://www.pymetrics.ai>

so that every form of responding can have a match to some kind of employment. Multidimensional measurement also allows us to optimize for prediction subject to fairness constraints (Wilson et al., 2021). Finally, the two-step approach lets pymetrics do selection, internal mobility, outplacement, a job marketplace, all based on one set of measures.

Thus, we have game-based psychological measurement with substantial plurality of purpose. To support this, the measures need to be robust, reliable, and as broadly as possible valid. The machine learning algorithm for selection models works to ensure that irrelevant or noisy measures aren't used to select people into a job, but even machine learning benefits from good "feature engineering" of its inputs. That is, using the tools of psychometrics, we audit measurement quality and find ways to improve it.

One of the metrics of quality we care about is reliability, or standard error of measurement, or signal to noise ratio. Those are all algebraic transformations of each other, and can be conditional or general depending on the circumstances. Internally at pymetrics, we talk about all three; this chapter discusses the issue primarily in terms of reliability.

Reliability is corroborating evidence for our validity argument. The preferred, primary evidence is a criterion validation study, but reliability helps us be sure that study is not capitalizing on chance. For our client contacts, many of whom have background in industrial-organizational psychology, reliability is a familiar concept and a reassuring metric to have available. In addition, reliability estimates help us evaluate proposed changes to measures, and help us make sense of the factor model we use for our explanations and descriptions. We have measures that legitimately measure non-redundant constructs, but we have also had some plain noisy ones, and reliability information helps us not to confuse the two. Notably, these measurements are delivered to machine learning models and not directly to humans, alleviating the need to communicate uncertainty in an individual score to an end user.

Games being games, our measure set is laden with eccentric scoring algorithms. To estimate reliability, we can't always rely on Cronbach's alpha, or get conditional standard errors from our latent trait estimation procedure. Some of our measures are strange beasts, and estimating their reliability is an adventure.

So this is a collection of short stories of reliability estimation in our applied context. I hope they prove encouraging or useful to somebody else who needs to be creative, or at the very least, they're entertaining.

These anecdotes include a case where we had interdependencies between the value of one measure and the reliability of another, and a case where we had an unusual scoring algorithm combined with extensively counterbalanced stimuli. There are two cases of composite scores, and finally, one case where the best answer came from process modeling.

2 Interdependencies: The Story of Easy or Hard

In modern test scoring, a measure's reliability may not be a uniform property applicable to all possible persons, scores and response patterns. We are accustomed to the standard error of measurement being higher at extreme scores, or perhaps for atypical response patterns. We can find the conditional standard error of measurement for a user, and then convert it to an individual reliability value applicable to a score; we can profile the quality of measurement across a population or across the range of measurement.

The situation becomes more complicated when the reliability of one measure depends in part on performance on *another* measure. This was the challenge presented by Easy or Hard.

Easy or Hard is adapted for brevity from the Effort-Expenditure for Rewards Task (EEfRT) (Treadway et al., 2009), a popular measure of motivation for rewards developed initially to study effort-based decision-making (Fig. 1). A psychometrics user is presented with a series of choices between an easy task and a hard task. Both tasks involve tapping a key or touch-screen hotspot a certain number of times within a certain interval, but the hard one is more demanding: sixty taps in twelve seconds instead of five times in three seconds. Each choice offers different game money rewards for the two tasks, with an equal probability of payout upon success, regardless of the task chosen. The probability of payout varies across trials. Easy or Hard has nine different expected value differences based on probability and amount of payout.

There is a total time limit of two minutes, after which the game ends. This makes Easy or Hard, in essence, a resource allocation task: a finite quantity of available time must be divided among task-choosing decision time, time to complete easy tasks, and time to complete hard tasks. Hard tasks have higher rewards, but can take up to twelve seconds to complete, and in addition to the random probability of no payout, it is realistically possible to fail the hard tap task. By contrast, it's essentially impossible to fail the easy task, but it is easy to complete it quickly and move on. The user is not required to wait out the three seconds. Two or three easy task payouts can easily exceed one hard task payout. Further, any time spent deciding between the tasks and their offered rewards is time without pay.

Some users scan the information and then pick a task based on calculation or heuristics. Some users always choose the easy task, and others always choose hard. The decision itself is a speeded task, as well. After five seconds, a task is assigned at random, which is undesirable as a strategy for obtaining rewards.

In addition to neuroscience findings linking hard task selection on the EEfRT to reward processing brain activity (Treadway et al., 2012), the EEfRT fits in a class of measures of extrinsic motivation (Ryan & Deci, 2000). In the workplace, a measure of extrinsic motivation can provide insight into how compensation may motivate employees, in particular when indicators of intrinsic motivation are also considered (Cerasoli et al., 2014). A multidimensional view of Easy or Hard begins to address this construct space.

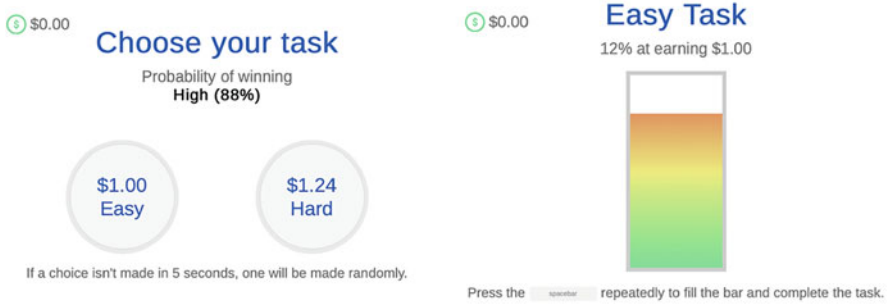


Fig. 1 *Easy or Hard.* *Left side: Task selection.* The user has five seconds to choose between an “Easy” and a “Hard” task. Either task, if completed successfully, gives the user a specified chance (here, 88%) of receiving a game-money reward. The reward is always larger for the “Hard” task. *Right side: Task completion.* A user selecting the “Easy” task must press the spacebar five times in three seconds in order to have the specified probability (here, 12%) of obtaining a game-money reward. A user selecting the “Hard” task must press the spacebar sixty times in twelve seconds in order to have the same chance of a higher reward

We are therefore, first, interested in a user’s probability of selecting the hard task under different expected value conditions, and overall. Some users show a consistent preference for the hard task even when the rewards are low and its selection is not strategic; this may be an indicator of one kind of intrinsic motivation. Another measure is the time taken to choose a task, an indicator of attention paid to the differential reward information.

The challenge in calculating the reliability of either the hard task preference measure or the response time measure is that both are conditional on the number of decision responses contributing to the measure, which is in turn conditional on the number of hard tasks selected (or assigned following timeout) and also the amount of time spent on the decision screen. If a user times out or nearly so, and always chooses or is assigned hard, that user can run out of total time in five tasks. We observe some cases of that in our data. We also see the occasional user that can race through thirty, fifty, or a hundred easy tasks, without reading anything. The content of a user’s choices determines the quality of that user’s measurement.

It was, however, possible to calculate Cronbach’s alpha for the subset of users that reached each item from the first to the 30th, which included fewer users for each additional item after five. We produced the whole table, but when we need a single number for reporting, we prefer the reliability for a user at the median number of trials.

As shown in Fig. 2, the quality of the RT measures improves rapidly in the first few trials, and more slowly thereafter for a long time. Averaging log (RT) is unsurprisingly better than averaging RT, but the same general behavior can be seen in both.

For the probability of choosing hard, however, the reliability dips for additional responses between about ten and fifteen. Around ten responses, the number of

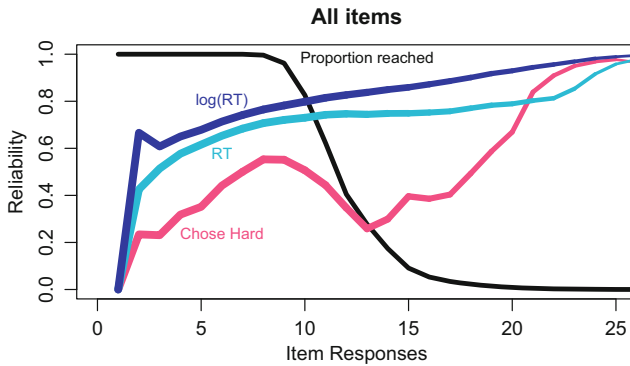


Fig. 2 *Easy or Hard: Reliability as a function of items reached.* Generally, measures of response time (averaged directly or as logarithms) increased in reliability with each additional task completed by the user. However, the reliability of preference for the “Hard” task (“Chose Hard”) dipped in reliability between ten and fifteen items reached. This is a range restriction phenomenon; due to the time limit, users can only reach items above the low teens by sometimes selecting “Easy” tasks. The higher the item count, the more “Easy” tasks are generally necessary to reach it. The range of possible scores on the preference measure is limited, and the variance reduced, on the right side of the plot. This in turn suppresses the reliability

users reaching each trial declines sharply, especially among the users with a high probability of picking hard. As a result, the variance of preference for hard across users drops, and by fifteen items, the complete-data sample is left with only users who chose several easy tasks. Even if the absolute error of measurement stays the same or improves, the reliability drops, relative to the surviving user population at each item. It’s a good reminder that reliability isn’t solely a property of the test, but of the composition of the sample.

3 Counterbalancing: The Story of Magnitudes

In working with tests that have subject matter constraints, curriculum coverage requirements, and other intentional non-uniformities in item content, there is a need to assess reliability in a way that recognizes that the mixture of content is intentional and would be consistent across hypothetical alternate forms. We frequently see, for example, a stratified version of Cronbach’s alpha in use.

What happens, though, when an intentional mixture of items meets a scoring method that isn’t a sum score, mean score, or latent trait score? This was the challenge we faced for two of the measures in Magnitudes.

Magnitudes is a pair of measures of Approximate Number Sense (ANS): Dots and Fractions (Fig. 3). Each measure is made up of quantitative comparison items. In Dots, the user selects the side that has the higher proportion of yellow dots. In Fractions, the user selects the larger fraction. They measure nearly the same

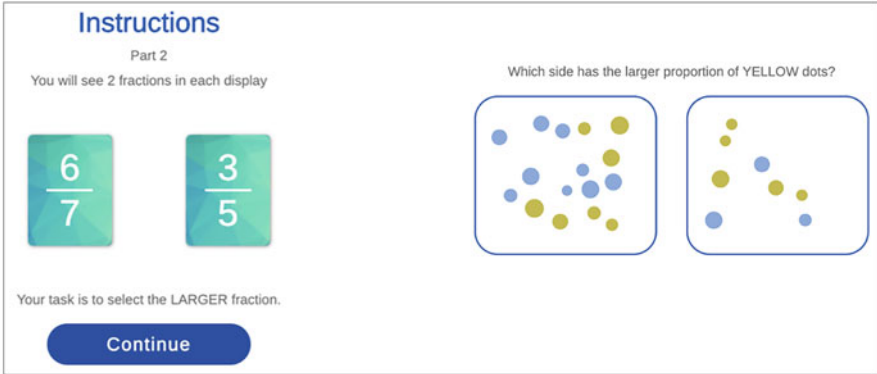


Fig. 3 *Magnitudes*. *Left side: Fractions*. A user is presented with two fractions side by side, and must choose the larger. The ratio of the two fractions is manipulated in order to increase or decrease difficulty. *Right side: Dots*. A user is presented two side-by-side arrays of mixed yellow and blue dots, and must select the array with proportionally more yellow dots. The numbers of blue and yellow dots are manipulated so as to contrive a range of ratios of proportions yellow. Automatic item generation is used to produce item clones with the specified dot counts, with the variably-sized dots distributed in a unique, non-overlapping pattern for each user

theoretical construct, but Fractions uses symbolic numbers and Dots uses non-symbolic quantity.

The primary metric of interest is the Weber fraction w for numerosity, which comes from psychophysics, and is the minimum difference between two quantities at which the greater quantity can be reliably recognized by an individual, expressed as a fraction of the smaller quantity. The Weber fraction w of an individual relates to the ratio r of the quantities compared, the improper fraction of the larger over the smaller quantity, which yields 75% accurate performance by that individual (Hunt, 2007):

$$w = r - 1 . \quad (1)$$

There is a model of performance implied by each of two theories of neural development; we currently use the Linear Spacing Model:

$$P(\text{correct}) = \Phi\left(\frac{1}{w} * \frac{(r - 1)}{\sqrt{r^2 + 1}}\right) \quad (2)$$

(Dietrich et al., 2016; Dehaene, 2007).

A separate w score is calculated for each section, because they measure different attributes (Baker & Thissen-Roe, 2021). Although the Weber fraction can be estimated as a latent trait (Thissen-Roe & Baker, 2021), we originally implemented it with estimation using a least squares algorithm (Price et al., 2012). We need to estimate the reliability of the least squares w as a baseline for evaluating the latent

trait version, as well as for the validity and documentation purposes common to all pymetrics measures.

For the least squares method, trials were grouped into four difficulty levels, ten trials each. The trials were also carefully counterbalanced for which side, left or right, has the correct response, and to foil heuristics applicable to each section, such as pixels of each color or fractions over one half. This counterbalancing creates an intentional mixture structure, similar to requirements for curriculum coverage. If we calculated reliability in a manner that failed to attend to the structure, the phenomena we have counterbalanced against (e.g., a preference for the left or right side in some users) would appear as a form of unreliability.

If we were using a sum or mean score, we could use stratified alpha. However, our scoring algorithm is a form of least squares regression. Previously in the literature, split-half reliability has been used with estimates of the Weber fraction for numerosity (Dietrich et al., 2016). Split-half reliability, with the Spearman-Brown prophecy formula for a full-scale estimate, works for any algorithm that can calculate a score from half the data, and have that score be an unbiased estimator of the full-scale score; specifically, for Spearman-Brown, the full-scale score is the unweighted sum or average of the half-scale scores. The former includes our least squares algorithm, and the latter is a linear approximation.

While stratified alpha is more commonly referenced in our discipline than stratified sampling of items for split-half, very often, the latter does work. We computed split-half reliability for the Weber fractions we obtained from each Magnitudes subtest, based on constrained halves that maintained the counterbalancing rules.

4 Composite Scores: The Story of Balloons

A relatively common challenge in estimating the reliability of scores used in employment testing and organizational development is that many of the scores, as used, are actually linear composites, or more complex functions, of multiple component measures. Composite scoring is done to reflect the task complexity and multifaceted nature of most jobs. For example, an employer may wish to allow candidates or employees to compensate for one area of weakness with another area of strength.

The pymetrics system has multiple layers of composite scoring (especially if taken broadly to include more functional forms than linear combinations). Not only are predictions of job fit made based on machine learning models using dozens of individual measures as inputs, some of those individual measures result from scoring functions that can be themselves interpreted as composites. Two types of these are described here; the first, simpler type includes several of the measures produced by the game Balloons.

Balloons is an implementation of the Balloon Analogue Risk Task (BART) (Fig. 4), which is used to measure risk-taking behavior (Lejuez et al., 2002; Lauriola

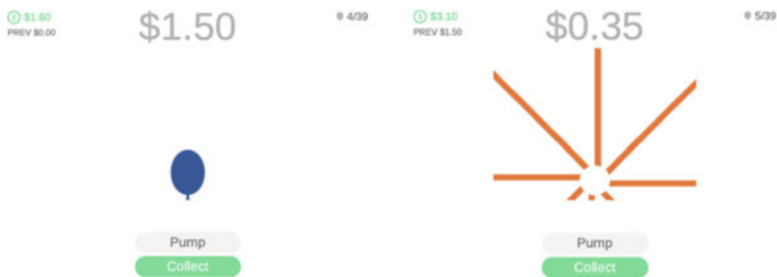


Fig. 4 *Balloons*. *Left*: Pumping up a blue balloon. This balloon is slowly expanding to fill the white space after thirty pumps, each good for \$0.05. *Right*: Oops! This orange balloon popped after seven pumps. The user will *not* receive \$0.35

et al., 2014; Charness et al., 2013). It taps into risk propensity and learning under uncertainty.

As a user, you pump up a balloon as much as you can before it pops. If you stop and it hasn't popped, you get paid game money for how much you pumped it up. If you pop the balloon, you get nothing. The popping happens at random within a specified distribution, not at the same number of pumps each time. There are three colors with different pop distributions. Each user is presented thirteen balloons of each color, giving the opportunity to learn over time.

High risk propensity in the BART has been found to be significantly associated with numerous real-world behaviors, including effective workplace maverickism (Gardiner & Jackson, 2012), propensity for gambling (Lejuez et al., 2002, 2003), and novel task learning speed (Humphreys et al., 2013). Low risk propensity, that is, caution, has been associated with more conservative workplace behavior and higher positions of power (Maner et al., 2007). Different job functions and types call for different levels of risk propensity (Nicholson et al., 2005), yet risk propensity remains stable within individuals (Josef et al., 2016). This combination makes it useful to pymetrics for assessing job fit.

The pymetrics implementation of *Balloons* produces several measures of change between the first and last few trials. Do you become more confident, and increase the number of pumps before you collect your money? Do you figure out where you need to stop to avoid popping so many balloons? Do you get faster or slow down at pumping each balloon? These change scores are computed as differences between ordinary mean scores. This, then, begins our discussion of scores that are derived from multiple simpler scores: composite scores.

Fortunately for the case of *Balloons*, reliability of difference scores is a solved problem. There is a straightforward formula for the calculation of reliability for difference scores, as given by Feldt and Brennan (1989):

$$\rho_{DD'} = 1 - \frac{\sigma_{X_2}^2 * (1 - \rho_{X_2X_2'}) + \sigma_{X_1}^2 * (1 - \rho_{X_1X_1'})}{\sigma_{(X_2-X_1)}^2} \quad (3)$$

With this formula, the reliability $\rho_{DD'}$ of the unit-weighted difference score $D = X_2 - X_1$ can be computed from the reliabilities $\rho_{XX'}$ and variances σ_X^2 of each component measure X , as well as the variance $\sigma_{(X_2 - X_1)}^2$ of the difference score itself. The latter works to index how non-redundant the two measures are; taking the difference of two measures that are too closely correlated will result in a low-variance, low-reliability, not-very-useful difference score.

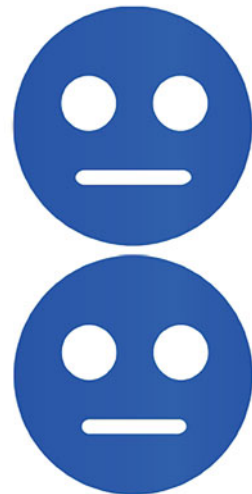
We were able to use this formula for changes in pump speed and pop frequency, and the establishment of a margin before a user expects a balloon to pop, as well as several other difference scores across the pymetrics games. Similar formulas exist, and are presented in the same chapter (Feldt & Brennan, 1989) for other common linear composites such as unweighted or weighted sums of two or more scores.

5 Composite Scores: The Story of Lengths

Not all derived scores, however, are linear composites. Estimates of reliability for derived scores were needed, but in a different way, for the game called Lengths.

Lengths is adapted from the Probabilistic Reward Task (PRT), which was developed to measure reward responsiveness and reinforcement learning (Pizzagalli et al., 2008, 2005). In Lengths, a user is asked to distinguish between two slightly different emoji faces, presented sequentially over seventy trials (Fig. 5). Both faces are variations on the still-face emoji, presented in white lines on a dark blue background for contrast. In the “long mouth” face, the flat-line mouth is a few pixels longer than in the “short mouth” face. The difference between faces is subtle, designed to be a Just Noticeable Difference under reasonable play conditions.

Fig. 5 *Lengths*. Long and short mouth face, enlarged



Users receive intermittent rewards for correct identification, but the reward schemes are different for the two faces, and one has a higher expected value; it is better for earning game money. The higher expected value emoji face is denoted “rich” and the lower expected value condition is denoted “lean.” There are two alternate forms with different reward patterns associated with each face, as well as different sequences of “long mouth” and “short mouth” emoji faces.

The primary measure of Lengths is a measure of reward sensitivity, operationalized as response bias. Do you favor the high-reward face when you’re not sure which one you’re looking at? Do you default to it? And by how much?

The formula for computation of response bias is adapted from earlier literature. Pizzagalli et al. (2008) used the following formula:

$$b = 0.5 * \log\left(\frac{C_{rich} * I_{lean}}{C_{lean} * I_{rich}}\right) \quad (4)$$

In this formula, the C and I elements are counts of correct and incorrect identifications in each condition. The Pizzagalli formula is equivalent to a difference of log odds ratios:

$$b = 0.5 * \left(\log\left(\frac{C_{rich}}{I_{rich}}\right) - \log\left(\frac{C_{lean}}{I_{lean}}\right)\right) \quad (5)$$

However, in rare cases, this formula can lead to division by zero or taking the logarithm of zero. In order to prevent such cases in our large operational setting, as well as to more conservatively score extreme observed performance (e.g. “lucky streaks” or forms of inattentive responding), each observed response count was increased by one-half response, a simple projection of what might have occurred if the game were extended. (A more sophisticated approach might have been the use of a Bayesian prior; however, the practical effect would be similar.)

$$b = 0.5 * \log\left(\frac{(C_{rich} + 0.5) * (I_{lean} + 0.5)}{(C_{lean} + 0.5) * (I_{rich} + 0.5)}\right) \quad (6)$$

There remains a simpler formulation as a difference of log odds ratios, albeit with the half-response adjustment included.

$$b = 0.5 * \left(\log\left(\frac{(C_{rich} + 0.5)}{(I_{rich} + 0.5)}\right) - \log\left(\frac{(C_{lean} + 0.5)}{(I_{lean} + 0.5)}\right)\right) \quad (7)$$

It is worth noting that these scoring formulae give relative emphasis to the less-common incorrect identifications. (They are less common in that individuals perform above chance; they are not rare.) Theoretically speaking, errors are more informative than correct responses.

As far as calculating reliability goes, the response bias measure is derived from two simple sum scores and two transformations of the same sum scores; the correct and incorrect counts for each face must add up to the number of times the face

is presented. The reliability of each sum score can be estimated simply. However, getting from those component reliabilities to the reliability of b , the response bias measure, is more complex.

Feldt and Brennan (1989) didn't cover products, or quotients, or logarithms of scores! And for good reason. Only linear composites have consistent reliability across users and scores, even if all the components behave according to classical test theory. In a multiplicative composite, the standard error of measurement contribution for each component score depends on the value of the other component score. There are similarly nonlinear behaviors in the standard error of measurement of proportions, exponents and logarithms.

However, as Feldt and Brennan (1989) did, we can apply standard propagation of error rules to obtain each individual's standard error of measurement in a large and representative sample of our user population. That standard error varies for different individuals. Therefore, as in other cases where conditional standard error of measurement is appropriate, we can calculate the marginal reliability across the population.

Error propagation turned out to be simplest for a weighted sum of two log odds components, although there were equally valid possibilities for calculating composite reliability using the original log-of-quotient-of-products version. Because the frequency of correct responses and the frequency of incorrect responses are linearly dependent—they sum to one—each log odds component is an expression of the form described by y in the equation below, where x is proportion incorrect and c is the half-response.

$$y = \log\left(\frac{1 - x + c}{x + c}\right) \quad (8)$$

From this, we can derive a function for the standard error of y in terms of the value of x and the standard error of x .

$$s_y = s_x * \left(\frac{1}{1 - x + c} + \frac{1}{x + c}\right) \quad (9)$$

Once the standard errors of the log odds components are obtained, they can be used with ordinary linear composite rules, as in the previous section, to produce a reliability estimate across the two mouth lengths.

Using this method, we obtained, as we had hoped, similar marginal reliability estimates for the two alternate forms, supporting their use in parallel.

6 Simulation and Modeling: The Story of Digits

When no formula seems immediately appropriate, we can sometimes gain insight into the reliability of a measure through process modeling. As an example, here is the story of Digits.

Digits is a visual forward digit span task. Several digits are presented on screen one at a time, after which the user is asked to type in those digits, in order. If the user gets it right, the task is repeated with a sequence one digit longer. If the user gets a sequence wrong, the next sequence has one less digit the next time. On the third error, the game terminates, and the primary score is the length of the longest sequence correctly recalled.

A wide variety of jobs call for a good working memory, the ability to hold information briefly in mind without external assistance. Importantly to pyometrics, memory span tasks have been found to show smaller differences between demographic groups than more general measures of cognitive ability (Verive & McDaniel, 1996); it is thus more feasible to create a fair composite score that includes digit span, when it is demonstrated to be job-relevant.

Digit span measures have been around for more than a century, and there are numerous established methods in the literature for computing their reliability, each dependent on the specifics of the task. Non-adapting versions of the digit span task (and other related memory span tasks) are amenable to estimation of split-half reliability or Cronbach's alpha (Waters & Caplan, 2003). Sometimes digit span reliability is obtained just by having some or all users do the task twice, and then correlating the results. However, in an effort to respect our users' time, we use a simple adaptive form and only present the task once.

In order to explore how error of measurement arises and manifests in Digits, we simulated user behavior using a simple item response model that has an error term built in. The model was a two-parameter normal ogive model, with a mean drawn from a distribution that approximates our observed scores, and a constant standard deviation across simulees, spans and trials.

While the standard deviation of the digit span score was observable in our data, the standard deviation parameter of the item characteristic curve (a transformation of item discriminability) was not directly observable. The simulation setup allowed us to test a range of values for the parameter, and map the effects on observable individual and population metrics. One such metric was the difference in length between the highest span with a correct response and the last span tested, which by definition got an incorrect response. The possible values are -1 , 0 and 1 , except in the rare case where a user enters no correct values, usually due to technical problems.

The objective of the simulation, then, was to relate the standard deviation parameter of the item characteristic curve to the proportions of each of the three possible last span deltas, and from that function, map the actual observed proportions back to a standard deviation parameter, or at least a small range of plausible parameter values. The standard deviation parameter can be translated back into a reliability estimate, using the concept that a reliability $\rho_{XX'}$ is the proportion of score variance σ_X^2 not attributable to error of measurement.

$$\rho_{XX'} = \frac{(\sigma_X^2 - \epsilon^2)}{\sigma_X^2} \quad (10)$$

Because the stopping rule depends on a single (third) error, the reliability of the measure is effectively the reliability of a single item response.

The assumption that the standard deviation parameter (or item discriminability) holds constant across all individuals and sequence lengths is important. There is no reason from the theory of working memory to believe it must be constant; however, it is necessary to obtain a constant estimate of reliability across all users and scores on Digits. Therefore, the simple model was used.

The form of the simulation study allowed for different approaches to estimating the standard deviation parameter. Optimization methods such as maximum likelihood could be used to obtain a single best value. First, however, we chose to plot the simulation results, in order to visually assess the range of plausible values, as well as checking to see that the model was plausible at all given the actual triplet of observed last span delta frequencies.

The simulation results, and the observed user proportions, are plotted in Fig. 6. If the model were perfect, the top, middle and bottom pairs of solid and dashed lines should cross at the same left-right location. They don't. That's a sign of model misfit. In particular, the model doesn't account for the handful of users with technical problems, and also under-predicts a small percentage of users that appear to give up at some point, whether due to frustration or interruptions. These few users have all three sequential errors right at the end, and often the kind of error that suggests not trying (e.g., blank, repetitive or very brief responses).

The misfit isn't severe. There is a plausible range of values defined by the places where the three pairs of lines do cross, and a modest amount of difference in the observed and predicted probabilities through most of that range. That whole

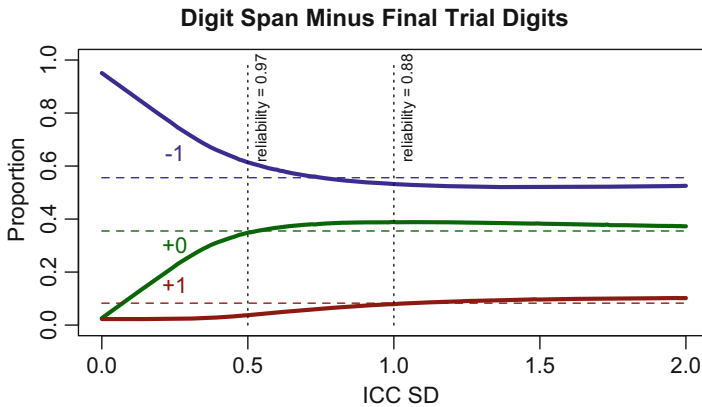


Fig. 6 *Digits*: Difference between two span scores as a function of the standard deviation of the item characteristic curve of recalling an individual sequence of digits. Solid lines represent the simulated frequencies of each last span delta according to the standard deviation parameter of the item characteristic curve for a single trial of recalling a digit sequence. Dashed horizontal lines represent observed frequencies in a population of actual users. Dashed vertical lines show a region of plausible values for the parameter, with the corresponding measure reliability noted

plausible range is consistent with values obtained in prior literature for the reliability of similar digit span tasks (Waters & Caplan, 2003). On the other hand, out toward the right of the plot, all of the curves flatten out, and don't change much. This suggests that neither eyeballing nor optimization is ever going to come to a highly certain and replicable result here. Nor is it likely to be apparent whether conditional standard error of measurement is needed.

For now, we choose to use the most conservative estimate of reliability in the plausible range, with a standard deviation of 1 and a reliability of 0.88.

7 Conclusion

As technology allows measurement to grow more complicated, integrated, and comprehensive, reliability becomes more difficult to estimate, but not less relevant. As with other forms of technical quality assurance, we must innovate to keep up with innovation in scoring, and rely on creativity to keep up with creativity in measurement. I hope these anecdotes provide some inspiration to you, my readers, when your own measures are up to their own shenanigans. Good luck to all of you!

Acknowledgments I would like to thank Frida Polli, founder of pymetrics. I would also like to thank Lewis Baker and Jackson Dolphin for the stories behind the games.

References

- Baker, L., & Thissen-Roe, A. (2021). Differences in symbolic and non-symbolic measures of approximate number sense. In M. Wiberg et al. (Eds.), *Quantitative psychology* (pp. x-x). New York: Springer.
- Cerasoli, C. P., Nicklin, J. M., & Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin*, *140*(4), 980–1008.
- Charness, G., Gneezy, U., & Imas, A. (2013). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior and Organization*, *87*, 43–51.
- Dehaene, S. (2007). Symbols and quantities in parietal cortex: Elements of a mathematical theory of number representation and manipulation. In P. Haggard & Y. Rossetti (Eds.), *Attention and Performance XXII Sensorimotor Foundations of Higher Cognition* (pp. 527–574). Cambridge: Harvard University Press.
- Dietrich, J. F., Huber, S., Klein, E., Willmes, K., Pixner, S., & Moeller, K. (2016). A systematic investigation of accuracy and response time based measures used to index accuracy. *PLoS One*, *11*(9), e0163076.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 105–146). New York: American Council on Education/Macmillan.
- Gardiner, E., & Jackson, C. J. (2012). Workplace mavericks: How personality and risk-taking propensity predicts maverickism. *British Journal of Psychology*, *103*(4), 497–519.
- Humphreys, K. L., Lee, S. S., & Tottenham, N. (2013). Not all risk taking behavior is bad: Associative sensitivity predicts learning during risk taking among high sensation seekers. *Personality and Individual Differences*, *54*(6), 709–710.

- Hunt, E. (2007). *The mathematics of behavior*. New York, NY: Cambridge University Press.
- Josef, A. K., Richter, D., Samanez-Larkin, G. R., Wagner, G. G., Hertwig, R., & Mata, R. (2016). Stability and change in risk-taking propensity across the adult lifespan. *Journal of Personality and Social Psychology*, *111*(3), 430–450.
- Lauriola, M., Panno, A., Levin, I. P., & Lejuez, C.W. (2014). Individual differences in risky decision making: A meta-analysis of sensation seeking and impulsivity with the balloon analogue risk task. *Journal of Behavioral Decision Making*, *27*(1), 20–36.
- Lejuez, C.W., Aklin, W. M., Zvolensky, M. J., & Pedulla, C. M. (2003). Evaluation of the balloon analogue risk task (bart) as a predictor of adolescent real-world risk-taking behaviours. *Journal of adolescence*, *26*(4), 475–479.
- Lejuez, C.W., Read, J. P., Kahler, C.W., Richards, J. B., Ramsey, S. E., Stuart, G. L., . . . Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: the balloon analogue risk task (bart). *Journal of Experimental Psychology: Applied*, *8*(2), 75–84.
- Maner, J. K., Gailliot, M. T., Butz, D. A., & Peruche, B. M. (2007). Power, risk, and the status quo: Does power promote riskier or more conservative decision making? *Personality and Social Psychology Bulletin*, *33*(4), 451–462.
- Nicholson, N., Soane, E., Fenton-O’Creevy, M., & Willman, P. (2005). Personality and domain-specific risk taking. *Journal of Risk Research*, *8*(2), 157–176.
- Pizzagalli, D. A., Iosifescu, D., Hallett, L. A., Ratner, K. G., & Fava, M. (2008). Reduced hedonic capacity in major depressive disorder: Evidence from a probabilistic reward task. *Journal of Psychiatric Research*, *43*(1), 76–87.
- Pizzagalli, D. A., Jahn, A. L., & O’Shea, J. P. (2005). Toward an objective characterization of an anhedonic phenotype: a signal-detection approach. *Biological Psychiatry*, *57*(4), 319–327.
- Price, G. R., Palmer, D., Battista, C., & Ansari, D. (2012). Nonsymbolic numerical magnitude comparison: Reliability and validity of different task variants and outcome measures, and their relationship to arithmetic achievement in adults. *Acta Psychologica*, *140*, 50–57.
- Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*, 54–67.
- Thissen-Roe, A., & Baker, L. (2021). Estimating approximate number sense (ANS) acuity. In M. Wiberg et al. (Eds.), *Quantitative psychology* (pp. x–x). New York: Springer.
- Treadway, M. T., Buckholtz, J. W., Cowan, R. L., Woodward, N. D., Li, R., Ansari, M. S., & Zald, D. H. (2012). Dopaminergic mechanisms of individual differences in human effort-based decision-making. *Journal of Neuroscience*, *32*(18), 6170–6176.
- Treadway, M. T., Buckholtz, J. W., Schwartzman, A. N., Lambert, W. E., & Zald, D. H. (2009). Worth the ‘EEFRT’? The effort expenditure for rewards task as an objective measure of motivation and anhedonia. *PloS One*, *4*(8), 1–9.
- Verive, J. M., & McDaniel, M. A. (1996). Short-term memory tests in personnel selection: Low adverse impact and high validity. *Intelligence*, *23*, 15–32.
- Waters, G. S., & Caplan, D. (2003). The reliability and stability of verbal working memory measures. *Behavior Research Methods, Instruments and Computers*, *35*(4), 550–564.
- Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., . . . Polli, F. (2021). Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Chicago, IL (pp. 666–677).

Detecting Testlet Effects in Cognitive Diagnosis Models



Youn Seon Lim

Abstract A testlet is a cluster of items that shares a common stimulus (e.g., a set of questions all related to the same text passage). The testlet effect calls into question one of the key statistical assumptions of any tests: local independence of the test item responses. Local dependence among test item responses is typically induced by the under-specification of the latent ability dimensions supposed to underlie a test. Hence, evaluating whether local independence holds for the items of a given test can be used as a diagnostic tool for detecting testlet effects. This study studied and compared the MH statistic, the Chi-squared statistic and the absolute deviations of observed and predicted corrections in detecting testlet effects in cognitively diagnostic tests. Various simulation studies were conducted to evaluate their performance under a wide variety of conditions.

Keywords Cognitive diagnosis models · Testlet effects · Mantel-Haenszel statistic · Chi-squared statistic · Absolute deviations of observed and predicted corrections

1 Introduction

A testlet is “a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow” (Wainer & Kiely, 1987, p. 190). A typical example is a reading comprehension test in which a reading passage is used as the stimulus for more than one item to measure examinees’ ability to comprehend the reading passage. Another example refers to ordering sentences to make a complete passages, where the items (sentences) are embedded in the passage itself. Responses to items within a testlet calls into question of the key statistical assumptions of any test: local independence

Y. S. Lim (✉)

Quantitative and Mixed Methods Research Methodologies, Educational Studies, University of Cincinnati, Cincinnati, OH, USA

e-mail: limyo@ucmail.uc.edu

of the test item responses. Local dependence among test items is typically induced by the under-specification of the measured latent dimensions by a test (i.e., Lim & Drasgow, 2019a,b; Rupp et al., 2010).

Various methods have been suggested for examining local dependence in cognitive diagnosis models. For example, de la Torre and Douglas (2004) evaluated item pair dependence using bivariate information. Templin and Henson (2006) expanded their method by using a parametric bootstrap method to estimate the distribution of the item association measures to estimate a p -value for their test statistic. Chen et al. (2013) used comparing the residual between the observed and expected Fisher-transformed correlation, and the residual between the observed and expected between log-odds ratios for the measure of association for each pair of items. Lim and Drasgow (2019a,b) also modified the Mantel Haenszel (MH) statistic for the measure. Those statistics have been used for the model fit or Q -matrix fit evaluation. This study evaluates the performances of the MH statistic (Lim & Drasgow, 2019a,b), the Chi-squared statistic $x_{jj'}$ (Chen & Thissen, 1997), the absolute deviations of observed and predicted corrections $r_{jj'}$ (Chen et al., 2013) in detecting testlet effects in cognitively diagnostic test in various simulation conditions.

2 Cognitive Diagnosis Models

Three cognitive diagnosis models were considered in this study: Deterministic-Input, Noisy “And” gate (DINA) model, Generalized Deterministic Inputs, Noisy “And” gate (G-DINA) model (saturated model), and Additive Cognitive Diagnosis Model (A-CDM).

Let Y_{ij} denote the binary item response of the i th examinee to the j th item, $i = 1, \dots, I$, $j = 1, \dots, J$ with 1 = correct and 0 = incorrect. Cognitive diagnosis models formulate the conditional distribution of item responses Y_{ij} given examinee latent attributes $\alpha_i = \{\alpha_{ik}\}$, for $k = 1, \dots, K$. (e.g., de la Torre & Douglas, 2004). Each entry α_{ik} indicates whether the i th examinee has mastered the k th attribute with 1 = mastered and 0 = not mastered. The binary $J \times K$ Q -matrix is an essential component of cognitive diagnosis models. The Q -matrix has a row for each item, $j = 1, \dots, J$, and a column for each attribute, $k = 1, \dots, K$. Each entry q_{jk} in the matrix indicates whether the k th attribute is required for the solution of the j th item with 1 = required and 0 = not required.

A common cognitive diagnosis model is the DINA model (e.g., Junker & Sijtsma, 2001). In this model, an ideal response η_{ij} is used to indicate whether all required attributes for the j th item are mastered by the i th examinee. The item response function (IRF) for the DINA model is

$$P(Y_{ij} = 1 \mid \alpha_i, s_j, g_j) = (1 - s_j)^{\eta_{ij}} g_j^{(1 - \eta_{ij})},$$

where $s_j = P(Y_j = 0 \mid \eta_j = 1)$ and $g_j = P(Y_j = 1 \mid \eta_j = 0)$.

Henson et al. (2009) proposed the Log-Linear Cognitive Diagnosis Model (LCDM). The LCDM can fit a full continuum of cognitive diagnosis models that range from fully compensatory models to fully conjunctive models. The DINA model can be written as a special case of the LCDM. In particular, if an item requires two attributes, the IRF can be written as

$$P(Y_{ij} = 1 \mid \alpha_i) = \frac{\exp[\lambda_j \alpha_1 + \lambda_j \alpha_2 + (\lambda_j) \alpha_1 \alpha_2 - \eta_j]}{1 + \exp[\lambda_j \alpha_1 + \lambda_j \alpha_2 + (\lambda_j) \alpha_1 \alpha_2 - \eta_j]},$$

where $\eta_j = -\ln(g_j/1 - g_j)$ and $\lambda_j = \eta_j + \ln(s_j - 1/s_j)$.

de la Torre's Generalized DINA (G-DINA) model is another example (de la Torre, 2011). Similar to the LCDM, the G-DINA model can be reduced to special cases of general cognitive diagnosis models with different link functions: identity, logit, and log. The framework of the G-DINA is based on the DINA model. However, the 2^K latent class memberships of the DINA model are partitioned into $2^{K_j^*}$ latent groups, where $K_j^* = \sum_{k=1}^K q_{jk}$ denotes the number of required attributes for item j . Let α_{ij}^* be the reduced attribute vector whose elements are the required attributes for item j . Then the probability that a test taker mastering the attribute pattern α_{ij}^* (i.e., all elements of α_{ij}^* would answer item j correctly) is given by

$$\begin{aligned} P(\alpha_{ij}^*) &= P(Y_{ij} = 1 \mid \alpha_{ij}^*) \\ &= \lambda_{j0} + \sum_{k=1}^{K_j^*} \lambda_{jk} \alpha_{ljk} + \sum_{k' > k}^{K_j^*} \sum_{k=1}^{K_j^* - 1} \lambda_{jkk'} \alpha_{ljk} \alpha_{lj k'} \dots + \lambda_{j1, \dots, K_j^*} \sum_{k=1}^{K_j^*} \alpha_{ljk}, \end{aligned}$$

where λ_{j0} is the intercept, λ_{jk} is the main effect, $\lambda_{jkk'}$ and $\lambda_{j1, \dots, K_j^*}$ are interaction effects.

Without the interaction terms, the G-DINA model becomes the Additive-CDM (A-CDM). The A-CDM is one of several reduced models that can be derived from the saturated G-DINA model. The IRF of the additive model is given by

$$P(\alpha_{ij}^*) = P(Y_{ij} = 1 \mid \alpha_{ij}^*) = \lambda_{j0} + \sum_{k=1}^{K_j^*} \lambda_{jk} \alpha_{ljk}.$$

Item j has $K_j^* + 1$ parameters in this model. The mastery of an attribute has a constant and direct impact on the probability of a correct response.

3 Fit Statistics for Local Dependence

Three fit statistics were evaluated in this study: MH statistic (Lim & Drasgow, 2019a,b), the Chi-squared statistic denoted by $x_{jj'}$ (Chen & Thissen, 1997), the absolute deviations of observed and predicted corrections denoted by $r_{jj'}$ (Chen et al., 2013).

Lim and Drasgow (2019a,b) modified the MH chi-square statistic which were originally introduced by Mantel and Haenszel (1959) to test for conditional independence of two dichotomous or categorical item responses j and j' by forming the row-by-column contingency table, conditional on the levels of the control variable C , where $c = 1, 2, \dots, 2^K = C$ proficiency class membership. Let $\{i_{j,j'c}\}$ denote the frequencies of examinees in the $2 \times 2 \times C$ contingency table, and then

$$\text{MH}\chi^2 = \frac{[\sum_c (i_{11c} - \sum_c E(i_{11c}))]^2}{\sum_c \text{var}(i_{11c})},$$

where $E(i_{11c}) = i_{1+c}i_{+1c}/i_{++c}$ and $\text{var}(i_{11c}) = i_{0+c}i_{1+c}i_{+0c}i_{+1c}/i_{++c}^2(i_{++c} - 1)$.

The Chi-squared statistic $x_{jj'}$ (Chen & Thissen, 1997) is computed by forming the row-by-column contingency table,

$$\chi^2 = \sum_j \sum_{j'} \frac{(i_{jj'} - E(i_{jj'}))^2}{E(i_{jj'})},$$

where $E(i_{jj'}) = E_{pq} = N \int P_j(\theta)^p P_j(\theta)^q [1 - P_i(\theta)]^{(1-p)} [1 - P_j(\theta)]^{(1-q)} f(\theta) d\theta$, where $P_i(\theta)$ is the trace link for item j , $f(\theta)$ is the population distribution. For cognitive diagnosis models, $E(i_{jj'})$ is estimated by an examinee's posterior distributions (Robitzsch et al., 2020).

The absolute deviations of observed and predicted corrections $r_{jj'}$ (Chen et al., 2013) is calculated by

$$r_{jj'} = |\mathbf{Z}[\text{Corr}(\mathbf{Y}_j, \mathbf{Y}_{j'})] - \mathbf{Z}[\text{Corr}(\mathbf{Y}_j, \mathbf{Y}_{j'})]|,$$

where $\text{Corr}(\cdot)$ is the Pearson's product-moment correlation, $\mathbf{Z}(\cdot)$ is the Fisher's transformation.

4 Simulation Studies

To investigate the performance of the MH statistic, the Chi-squared statistic $x_{jj'}$, the absolute deviations of observed and predicted corrections $r_{jj'}$, a variety of simulation conditions were studied by crossing the numbers of examinees I , items J , and examinees' latent attribute distributions ρ for three different cognitive diagnosis models.

For each simulation condition, a set of item response vectors was simulated for 100 replications. Item response data of sample sizes $I = 500$ (small), or 2000 (large) were drawn from a discretized multivariate normal distribution $MVN(0_K, \Sigma)$, where the covariance matrix Σ has unit variance and common correlation $\rho = 0.3$ (low) or 0.6 (high). Test lengths $J = 20$ (short) or 40 (long) were studied. A Q -matrix was generated randomly from a discrete uniform distribution on the

Table 1 Correctly specified Q ($K = 3$)

Item	k_1	k_2	k_3
1	1	0	0
2	0	1	0
3	0	0	1
4	1	0	0
5	1	0	0
6	1	1	0
7	1	0	0
8	0	1	0
9	0	0	1
10	0	1	0
11	1	1	0
12	1	1	0
13	1	0	0
14	0	1	0
15	0	0	1
16	1	0	0
17	0	1	0
18	0	0	1
19	1	0	0
20	1	0	0

Table 2 T-Matrix: testlet specification ($M = 2$)

Testlet	Item																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
M_1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M_2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0

maximum $2^K - 1$ possible q-vectors for each condition and fixed for replications. The correctly specified Q -matrix for $J = 20$ is presented in Table 1. The Q -matrix for $J = 40$ was obtained by duplicating this matrix two times to study the longer length of item under the same attribute specification conditions.

Data were generated using three different models: the DINA model, A-CDM, and a saturated model (i.e., the 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 $P(\alpha_{ij}^*)_{\min} = 0.10$ and $P(\alpha_{ij}^*)_{\max} = 0.90$, where α_{ij}^* was the reduced attribute vector whose components are the required attributes for the j th item (see de la Torre, 2011, more details).

A fixed and pre-specified Item-by-testlet T -matrix was utilized to simulate testlet data. The entry t_{mj} of the T -matrix indicates whether the m th testlet, for $m = 1, 2, \dots, M$, includes the j th item. For each replication of 100 replications, the transpose of T -matrix shown in Table 2 was combined with Q -matrix ($K = 3$) in Table 1, to simulate item responses. A model was fitted only with the Q -matrix ($K = 3$).

The R Core Team (2020) was used for the estimation in this study (CDM package Robitzsch et al., 2020).

The MH statistic, Chi-squared statistic $x_{jj'}$ (Chen & 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. 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.

5 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. The type 1 errors and power rates of the three statistics are reasonable for detecting testlet effects in cognitive diagnosis models.

5.1 Type I Error Study

In this simulation study, the correctly specified Q -matrices ($K = 3$) were used to fit the data to examine type I error rates. The summarized rejection rates are reported in Table 3. The type I error rates of the $r_{jj'}$ became conservative when the numbers of items J and examinees I were increased. The Chi-squared test statistic $x_{jj'}$ was very conservative, with type I error rates below 0.024. The MH statistic got consistent under all conditions when item $J = 40$.

Table 3 Type I error study when $K = 3$

I	$J = 20$						$J = 40$					
	α with $\rho = 0.3$			α with $\rho = 0.6$			α with $\rho = 0.3$			α with $\rho = 0.6$		
	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$
<i>DINA model</i>												
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
<i>A-CDM</i>												
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
<i>Saturated model</i>												
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

Table 4 Simulation study: testlet dependent data

I	S	J = 20						J = 40					
		α with $\rho = 0.3$			α with $\rho = 0.6$			α with $\rho = 0.3$			α with $\rho = 0.6$		
		MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$	MH	$x_{ij'}$	$r_{ij'}$
<i>DINA model</i>													
500	T	0.928	0.925	0.946	0.803	0.869	0.925	0.815	0.853	0.900	0.896	0.903	0.937
	E	0.052	0.050	0.113	0.041	0.080	0.168	0.042	0.103	0.188	0.044	0.107	0.198
2000	T	0.999	0.998	0.999	0.968	0.995	1.000	0.983	0.993	0.996	0.977	0.996	0.998
	E	0.131	0.058	0.124	0.073	0.174	0.267	0.052	0.271	0.372	0.056	0.271	0.366
<i>A-CDM</i>													
500	T	0.713	0.770	0.848	0.672	0.779	0.862	0.733	0.758	0.844	0.683	0.769	0.858
	E	0.041	0.041	0.081	0.037	0.065	0.117	0.043	0.047	0.094	0.042	0.076	0.137
2000	T	0.998	0.999	1.000	0.987	0.995	0.999	0.996	0.995	0.998	0.989	0.993	0.996
	E	0.075	0.105	0.172	0.063	0.197	0.281	0.049	0.140	0.217	0.048	0.272	0.362
<i>Saturated model</i>													
500	T	0.535	0.606	0.708	0.448	0.562	0.702	0.567	0.608	0.714	0.456	0.577	0.691
	E	0.039	0.032	0.071	0.040	0.039	0.084	0.040	0.043	0.087	0.041	0.060	0.960
2000	T	0.922	0.963	0.978	0.845	0.928	0.955	0.945	0.953	0.979	0.874	0.929	0.960
	E	0.070	0.079	0.140	0.064	0.149	0.234	0.050	0.106	0.178	0.049	0.199	0.287

5.2 Power Study: Testlet Model

As shown in Table 4, high rejection rates for testlet dependent item pairs T were obtained for the three statistics (i.e., 0.803 or above in the DINA model, 0.672 or above in the A-CDM, and 0.448 or above in the saturated model). The power rates were moderately consistent under all conditions. Unlike the Chi-squared statistic $x_{jj'}$, and transformed correction statistic $r_{jj'}$, the rejection rates of the MH statistic for the item pairs in which only one item of a pair was testlet-dependent E were low (i.e., 0.075 or below). This implies that the MH test can play an important role in detecting only testlet dependent items. Not surprisingly, the performance of the test tends to slightly deteriorate in the saturated model.

6 Discussion

The simulation studies investigated the usefulness and sensitivity of the MH statistic, the Chi-squared statistic $x_{jj'}$, the absolute deviations of observed and predicted corrections $r_{jj'}$ in a variety of cognitive diagnosis modeling settings with testlet dependent items. The primary findings are that most type I error rates of the three different statistics were around the nominal significance level of 0.05. Furthermore the statistics perform reasonably well in detecting testlet dependent items. Nonetheless, the statistics are somewhat conservative and less sensitive to

different model settings. In summary, the statistics might be a promising tool for detecting testlet effects in cognitive diagnostic modeling.

For the popularity of testlets in large-scale assessments, it is necessary to investigate the issues related to testlet effects in cognitive diagnosis models. Ignoring testlet effects leads to inaccurate estimates of item parameters and misclassifications of examinees depending on the strengths of testlet effects with minimal influences of other properties of test constructions and administration (Lim et al., 2022). A few (unpublished) dissertations and two or three papers (e.g., Hansen, 2013) study testlet effects—but mainly in terms of how to model testlet effects. Till now, few testlet-effect detection procedures for cognitive diagnosis model have been investigated. Therefore, the significance of this study lies in investigating test statistics to detect testlet effects.

This study is not without limitations. One limitation is that the performance of the statistics was not evaluated with an empirical data. Another limitation is that the statistics were investigated with simple cognitive diagnosis models with testlets. With those limitations, researcher recommends further studies to be conducted with more complex cognitive diagnosis models and real datasets. Furthermore, the findings show that a cognitive diagnosis model that accounts for testlet effects is necessary.

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.
- de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333–353.
- Hansen, M. (2013). *Hierarchical Item Response Models for Cognitive Diagnosis*. Doctoral Dissertation, University of California, LA.
- Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log linear models with latent variables. *Psychometrika*, *74*, 191–210.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272.
- Lim, Y. S., & Drasgow, F. (2019a). Conditional independence and dimensionality of nonparametric cognitive diagnostic models: A test for model fit, *Journal of Classification*, *36*, 295–305.
- Lim, Y. S., & Drasgow, F. (2019b). Assessing the dimensionality of the latent attribute space in cognitive diagnosis through testing for conditional independence. In Wiberg, M., Culpepper, S., Janssen, R., González, J., & Molenaar, D. (Eds.). *Quantitative psychology research* (pp. 183–194). New York, NY: Springer.
- Lim, Y. S., Fangxing, & B., Kelcey B. (2022). Sensitivity of Cognitive Diagnosis Models to Local Item Dependence has been selected for presentation. In: *Annual Conference of the National Council on Measurement in Education (NCME) 2022*, San Diego, CA.
- 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.

- R Core Team (2020). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. <https://www.R-project.org/>.
- Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2020). *CDM: Cognitive diagnostic modeling*. R package version 3.4–21.
- Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic assessment: Theory, methods, and applications*. New York: Guilford.
- Templin, J. L., & Henson, R. A. (2006), Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287–305.
- Wainer, H & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case study for testlets. *Journal of Educational Measurement*, *24*, 195–201.

Comparison of Estimation Algorithms for Latent Dirichlet Allocation



Constanza Mardones-Segovia , Hye-Jeong Choi, Minju Hong ,
Jordan M. Wheeler , and Allan S. Cohen 

Abstract Latent Dirichlet Allocation (LDA; Blei et al., *J Mach Learn Res* 3:993–1022, 2003) is a probabilistic topic model that has been used to detect the latent structure of examinees' responses to constructed-response (CR) items. In general, LDA parameters are estimated using Gibbs sampling or variational expectation maximization (VEM). Relatively little evidence exists, however, regarding the accuracy of either algorithm in the context of educational research, such as small numbers of latent topics, small numbers of documents, short average lengths of documents, and small numbers of unique words. Thus, this simulation study evaluates and compares the accuracy of parameters estimates using Gibbs sampling and VEM in corpora typical of educational tests employing CR items. Simulated conditions include number of documents (300, 700, and 1000 documents), average answer length (20, 50, 100, and 180 words per document), vocabulary of unique words in a corpus (350 and 650 unique words), and number of latent topics (3, 4, 5, 6, and 7 topics). Accuracy of estimation was evaluated with root mean square error. Results indicate both Gibbs sampling and VEM recovered parameter estimates well but Gibbs sampling was more accurate when average text length was small.

Keywords Gibbs sampling · Variational expectation maximization · Latent Dirichlet allocation

C. Mardones-Segovia (✉)

University of Georgia, Department of Educational Psychology, Athens, GA, USA
e-mail: cam04214@uga.com

H.-J. Choi

The University of Georgia, Athens, GA, USA

The Human Resources Research Organization, Louisville, KY, USA

M. Hong · J. M. Wheeler · A. S. Cohen

University of Georgia, Athens, GA, USA

1 Introduction

Constructed response (CR) items have been widely used for measuring the reasoning and inquiry skills in examinees' answers on educational tests (Crossley et al., 2016). Typically, CR items are hand graded using a rubric. In large-scale settings, however, hand grading is expensive and time-consuming. Recently, due to the availability of algorithmic methods, such as Latent Semantic Analysis (Deerwester et al., 1990), it has been possible to mirror human grading of CR items, automating the scoring of a considerable number of answers in a fraction of the time needed for human grading.

More recently, another topic model, latent Dirichlet allocation (LDA; Blei et al., 2003), has been applied in the assessment setting to enhance the information gained from examinees' responses to CR items (e.g., Cardozo-Gaibisso et al., 2019; Kim et al., 2017). LDA is a probabilistic clustering model developed to analyze the latent thematic structure of collections of textual data. In the context of educational measurement, this type of topic model uses word co-occurrences to extract the latent topic structure in a collection of examinees' responses (Choi et al., 2019; Xiong et al., 2019). LDA parameters have been estimated using the variational expectation maximization algorithm (VEM; Blei et al., 2003) or Gibbs sampling (Griffiths & Steyvers, 2004). Relatively little evidence exists, however, regarding accuracy of the two algorithms in conditions common in educational research, such as small numbers of latent topics, small numbers of examinees, short lengths of answers, and small numbers of unique words.

Different researchers have pointed out that text length and the sample of documents influence the accuracy of LDA parameters. Short text lengths, for example, do not seem to contain the necessary word co-occurrences to estimate the parameters effectively (e.g., Hu et al., 2009). Further, the size of the corpus of documents seems to be the salient aspect that affects the performance of LDA as a small sample of documents is not sufficient to identify the latent topics (Tang et al., 2014).

Although applying LDA in answers to CR items has shown promising results, responses to test items tend to be smaller and shorter in length than the corpora for which LDA was initially developed. More importantly, the terms in the answers tend to be constrained by the prompts or questions. Thus, the vocabulary of unique words is also smaller than in the usual LDA data. Therefore, it is not clear whether the usual corpora of textual responses to CR items are enough to estimate LDA parameters accurately. This simulation study, then, evaluated the accuracy of parameter estimates using Gibbs sampling and VEM under conditions typical of educational tests.

2 Latent Dirichlet Allocation (LDA)

LDA is a probabilistic topic model designed to detect the latent structure in a corpus of textual documents. This cluster tool assumes that a corpus is a collection of D documents. Each document is a mixture over K topics, and each topic is a probability distribution over the vocabulary of V unique words (Blei et al., 2003). LDA estimates two structural parameters: topics and topic proportions. Topics or topic-word proportions are the proportion of assignment of topics over the vocabulary of unique words. Topic proportions or document-topic probabilities are the mixture of document's probabilities that are assign to each of K topic.

The main parameters in LDA are topics, also referred to as topic-word probabilities, and the topic proportions, also referred to as document-topic probabilities. Topics, denoted by ϕ_k , where $k \in 1, 2, \dots, K$, are a set of probabilities over the vocabulary in the corpus. Topic proportions, expressed by θ_d , where $d \in 1, \dots, D$, are the mixture of document proportions that are assigned to each K topic. Additionally, LDA estimates a topic membership for every word in each document. Topic assignments are denoted by $z_{d,n}$, where $d \in 1, \dots, D$, $n \in 1, \dots, N^{(d)}$, and $N^{(d)}$ indicates the number of words in document d .

For a given document d and topic k , the LDA model assumes the following distributions for the latent parameters: topics (ϕ_k) follow a Dirichlet distribution with a prior hyperparameter β , topic proportions (θ_d) follow a Dirichlet distribution with a prior hyperparameter α , and topic assignments ($z_{d,n}$) follow a multinomial distribution with parameters θ_d and $N^{(d)}$. Given the topic assignments for each word, LDA estimates the topics and topic proportions by assuming the words for each document ($w_{d,n}$) follow a multinomial distribution with parameters ϕ_k and $N^{(d)}$ (Ponweiser, 2012). The joint probability distribution of all latent and observed variables for LDA is expressed as:

$$p(w, z, \theta, \phi | \alpha, \beta) = \prod_{k=1}^K p(\phi_k | \beta) \prod_{d=1}^D p(\theta_d | \alpha) \prod_{n=1}^{N^{(d)}} p(z_{d,n} | \theta_d) p(w_{d,n} | \phi_k, z_{d,n}), \quad (1)$$

where $p(\phi_k | \beta)$ is the probability distribution of all topics over the vocabulary in the corpus, $p(\theta_d | \alpha)$ is the probability distribution of the topic proportions of a document, $p(z_{d,n} | \theta_d)$ is the probability distribution of the topic assignments of a document, and $p(w_{d,n} | \phi_k, z_{d,n})$ is the probability distribution of the observed words of a document (Ponweiser, 2012). Integrating over θ and ϕ , and summing over z , the marginal likelihood of a document is:

$$p(w | \alpha, \beta) = \int_{\phi} \int_{\theta} p(\theta | \alpha) p(\phi | \beta) \prod_{n=1}^N \sum_{z_n} p(z_n | \theta) p(w_n | z_n, \phi) d\theta d\phi \quad (2)$$

The marginal likelihood of each document is used to calculate the likelihood of a corpus by taking the product of each document’s marginal likelihood. It can be expressed as:

$$p(W|\alpha, \beta) = \prod_{d=1}^D p(w_d|\alpha, \beta) \quad (3)$$

Summing over the combination of all the topic assignments in Eq. 2, however, is computationally intractable. Thus, it is not possible to obtain exact estimates of the parameters (Blei et al., 2003).

2.1 Estimation Method

Different methods have been used to estimate LDA parameters, such as variational approximation and Gibbs sampling (Blei et al., 2003; Griffiths & Steyvers, 2004).

Gibbs Sampling

Gibbs Sampling algorithm approximate the multivariate probability distribution by randomly sampling a Markov chain of samples until they reach a stationary distribution (Taddy, 2012). Note that samples of the beginning of the chain tend to be correlated (Li et al., 2009). Thus, they may not be an accurate representation of the distribution. Therefore, samples of the beginning of the Markov chain are discarded and parameter estimates are obtained from the predictive posterior distribution. One of the drawbacks of this algorithm is that it requires a long convergence process to find the posterior of the document-topic distribution. As a result, Gibbs sampling is a slow and computationally time intensive estimation method (Kim, 2020).

Variational Expectation Maximization (VEM)

VEM is a deterministic algorithm to infer the posterior of the distribution of the latent variables. Instead of sampling the posterior, VEM uses optimization to approximate the lower bound of the log-likelihood. Thus, it is an estimation method that tends to be faster and computationally less time intensive than MCMC methods (Blei et al., 2017). To do this, VEM imposes a family of tractable simpler distributions over the latent variables of the LDA, i.e., to estimate the topic-proportions and topic assignments. VEM consists of two steps: the E-step, which find the optimizing values of the topic-proportion and topic assignment, and the M-step, which maximizes the lower bound of the log-likelihood by minimizing the Kullback-Liebler Divergence (KLD) between the approximation of the true and the estimated distribution (Taddy, 2012). It is important to note that although VEM is faster than the Gibbs sampling, it may underestimate the variance of the posterior (Blei et al., 2017).

What can affect the estimation procedure in LDA?

As Eq. 3 indicates, the Dirichlet prior hyperparameters play an important role in the estimation of the LDA model (Syed & Spruit, 2018). The Dirichlet distribution

assumes that parameters are mutually independent (Geiger & Heckerman, 1997). That is, the proportions of documents and words within topics are assumed to be independent. When these hyperparameters are larger than one, documents and words are equally likely to belong to all topics. Conversely, if the hyperparameters are less than one, it is more likely to produce fewer but more dominant topics (Syed & Spruit, 2018; Wallach et al., 2009).

The numbers of topics, documents, words, and text lengths have also been shown to affect the estimation of the LDA parameters. Tang et al. (2014) reported the accuracy of LDA models decreased considerably for a small number of documents when $\beta = 1$, regardless of average text length. Similarly, LDA performed poorly, when average text length was small, even for large numbers of documents. LDA performed better, however, when the distance between topics was larger. i.e., when topics were more distinguishable. Zuo et al. (2016), and Chen et al. (2016) also noted that LDA failed to provide accurate results when documents with small average text lengths contained low word counts. Further, Kwak (2019) suggested that the larger the vocabulary and the number of words per document, the more accurate the estimates of the topics and topic proportion, respectively. Kwak (2009) also reported that using hyperparameters of $\alpha = 50/K$ and $\beta = 200/V$ from Griffiths and Steyvers (2004), the accuracy of estimating ϕ_k appeared to depend on the number of unique words, whereas the performance of θ_d tended to be influenced by average text length.

3 Methods

3.1 Simulation Design

The aim of this simulation study was to investigate and compare the accuracy of LDA parameters estimated using Gibbs sampling and VEM in situations common in the assessment setting, i.e., small sets of documents, small numbers of unique words, short answer lengths, and small numbers of topics. Simulation conditions were chosen following the criteria mentioned in previous research on CR items (e.g., Hellman et al., 2020; Kim et al., 2017; Kwak, 2019; Xiong et al., 2019). Simulation conditions included the number of documents in a corpus (small = 300, medium = 700, and large = 1000 words), the number of unique words (small = 350, and 650 words), average length of documents (very small = 20 words, small = 50 words, medium = 100 words, and large = 180 words), and the numbers of topics in a corpus (3, 4, 5, 6, and 7 topics). Additionally, three sets of hyperparameters for the LDA models were chosen to emulate the effects on detection of the latent topic structure for distributions of documents and words typically found in measurement contexts (e.g., Kim et al., 2017; Wheeler et al., 2021; Xiong et al., 2021). Thus, this study used hyper-parameters that generate few but more distinguishable topics ($\alpha = 0.5 \& \beta = 0.05$, $\alpha = \frac{1}{K} \& \beta = 1$, $\alpha = \frac{1}{K} \& \beta = 0.5$). Due to the computational

cost of estimation of an LDA model (Wheeler et al., 2021), and that previous results have suggested that a great number of replications are not necessary to get precise results (Cohen et al., 2010), all conditions were crossed with 50 replications for each.

Documents, words, and topic assignments were simulated following the generative process described in Blei et al. (2003). For each LDA model, topics were drawn from a Dirichlet distribution using the hyperparameter for β . Similarly, for each document, topic proportions were drawn from a Dirichlet distribution using the hyperparameters above for α . The number of words in a document were drawn from a Poisson distribution given the number of documents and average text length. Next, for every word in a document, LDA estimates the topic assignment following a multinomial distribution with parameters θ_d and $N^{(d)}$. Then, a corpus of documents is sampled from a multinomial distribution given $z_{d,n}$ and ϕ_k . Finally, a document-term matrix is created to estimate candidate LDA models.

3.2 *Parameter Estimation*

LDA models were estimated using two algorithms: Gibbs sampling and VEM (Blei et al., 2003; Griffiths & Steyvers, 2004), both are implemented in the R package `topicmodels` (Grün & Hornik, 2011). As mentioned above, parameter estimates from Gibbs sampling were obtained from the predicted posterior distribution. In this study, 10,000 iterations were used as a burn-in and 5000 as post-burn-in to estimate the posterior. The same set of hyperparameters were used as priors to simulate the data. For VEM, LDA parameters were estimated using two procedures: (1) fixing α to the same hyperparameters used to simulate the data and (2) estimating α . The β hyperparameters were estimated freely as it cannot be fixed in the R package `topicmodels`.

3.3 *Label Switching*

Label switching may occur in LDA when the topics change their membership label either or both within iterations of a single MCMC chain or between chains. This study used the cosine similarity between simulated and estimated parameters to detect label switching. Cosine similarity values closer to one indicate that generated and estimated topics are similar, and that label switching had not occurred. Cosine similarity values close to zero or below were interpreted to mean that the simulated and estimated parameters were not similar, and therefore, label switching had occurred (Wheeler et al., 2021).

After correcting for label switching, the performance of Gibbs sampling and VEM were assessed for each condition in this study. The performance of an estimator can be measured by the bias and precision for estimating model parameters. Mean square error (MSE; Tietjen 1986), for example, calculates the mean of the

square differences by taking the variance of the estimates and the square mean error (Walther & Moore, 2005). Since MSE squares the differences, we used the root mean square error (RMSE) to have a measure in the same scale as the original data. RMSEs were estimated in Eq. 4, where R represent the number of replications, \hat{y}_r are the estimated parameters, and y_r are the true parameters. Values near zero suggest parameters estimate were more accurate.

$$RMSE = \sqrt{\sum_{r=1}^R \frac{(\hat{y}_r - y_r)^2}{R}} \tag{4}$$

4 Results

The accuracy of Gibbs sampling and VEM for estimating LDA parameters are presented in Figs. 1, 2 and 3. In general, estimation was faster and less computationally intense using VEM than Gibbs sampling. Both algorithms performed well at estimating topic-word proportions, regardless of the other conditions. The performance of the algorithms to estimate the document-topic proportions, however, appeared to decrease under certain conditions.

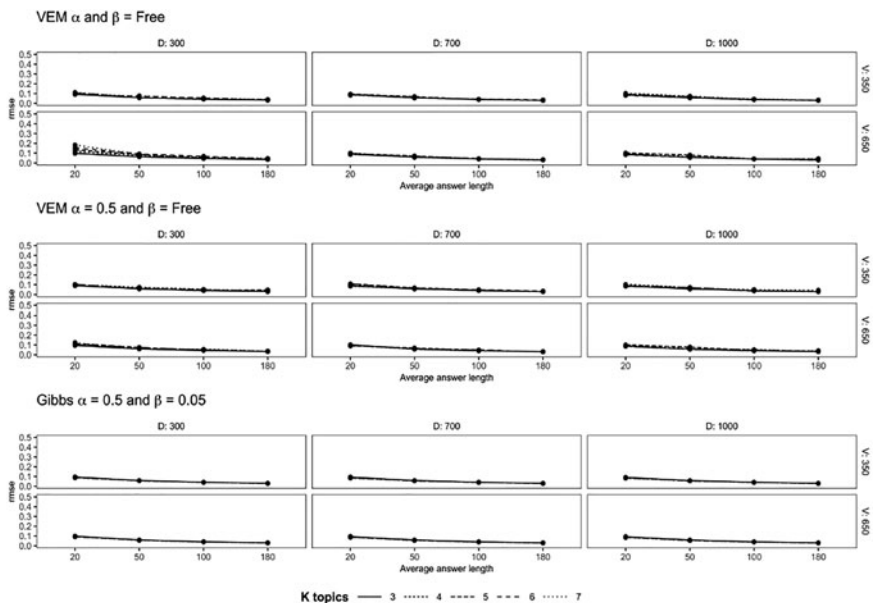


Fig. 1 Comparison of RMSE results for document -topic proportions when the true hyper-parameters were $\alpha = 0.5$ & $\beta = 0.05$

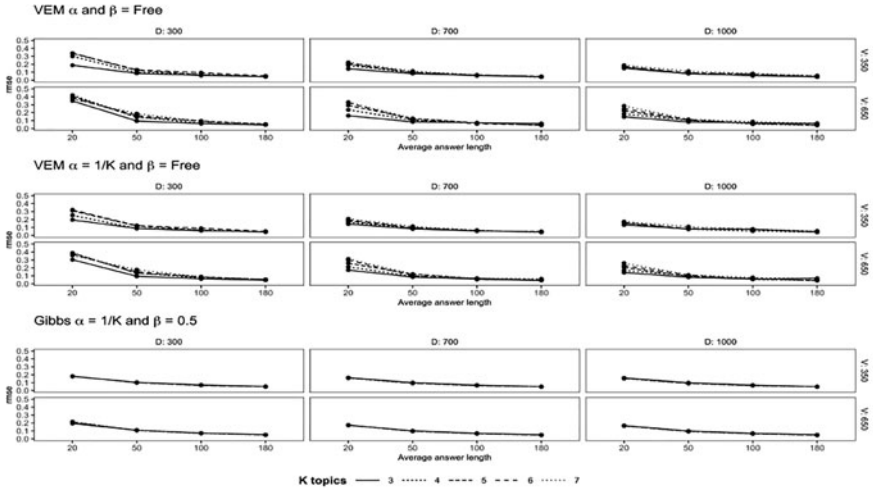


Fig. 2 Comparison of RMSE results for document -topic proportions when the true hyper-parameters were $\alpha = \frac{1}{K}$ & $\beta = 0.5$

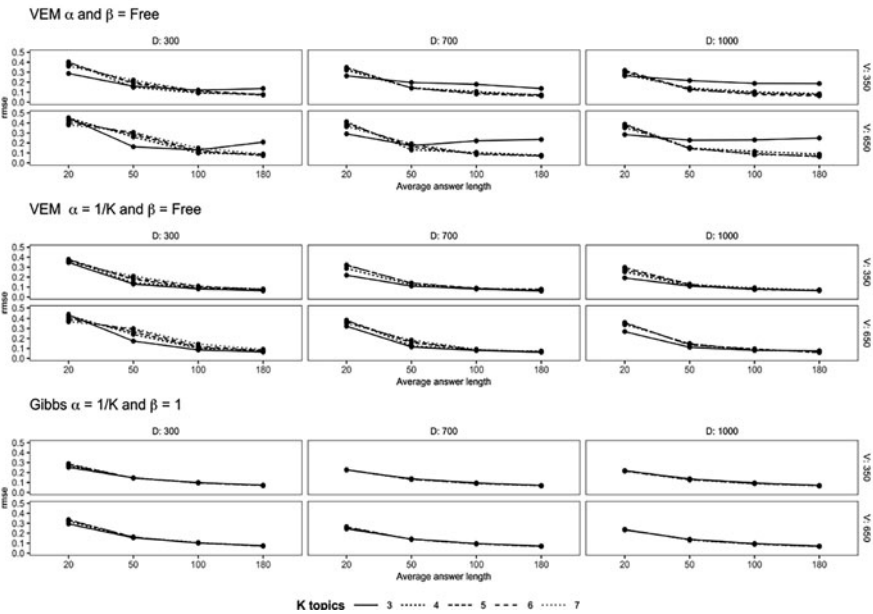


Fig. 3 Comparison of RMSE results for document -topic proportions when the true hyper-parameters were $\alpha = \frac{1}{K}$ & $\beta = 1$

The accuracy of estimating the topic proportion tended to be better across conditions when the hyperparameters used were $\alpha = 0.5$ & $\beta = 0.05$. Under this condition, RMSE values tended to be close to zero. No apparent differences were found between the performance of the Gibbs sampling and the VEM algorithm in estimating the document-topic proportion. The accuracy of both algorithms, however, tended to decrease when the set of hyperparameters were $\alpha = 1/K$ & $\beta = 0.5$, and $\alpha = 1/K$ & $\beta = 1$.

When the hyperparameters were $\alpha = 1/K$ & $\beta = 0.5$, Gibbs sampling and VEM seemed to be accurate in estimating topic proportions except when the average text length was 20 words. Under this condition, accuracy of estimating topic proportions tended to decrease, particularly when VEM was used to estimate the document-topic proportions with 300 documents.

Similarly, when the set of hyperparameters were $\alpha = 1/K$ & $\beta = 1$, the accuracy of the algorithms appeared to be influenced by the average text length. In general, Gibbs sampling was better for estimating topic proportions than VEM, particularly when average text length was less than 100 words and the size of the corpus was 300 documents. In addition, RMSE values tended to be higher when α and β were freely estimated.

5 Discussion

LDA has become popular for analyzing very large corpora of text. Recently, it has also been applied to analyze examinees' answers to CR items. Important differences in the assessment setting include often smaller numbers of documents, fewer unique words, and smaller average text length than the corpora typically used with LDA. In this study, the accuracy of Gibbs sampling and VEM algorithms were studied in estimating the LDA parameters in conditions that are usually seen in CR answers to test questions.

This study suggested that the topic-word distribution does not seem to depend on the estimator methods, or the conditions tested. Although some differences were found in parameter estimates using both algorithms, these differences were in the second decimal of RMSE values. Therefore, those discrepancies do not seem to be significant.

Results suggest some impact of estimation method on the accuracy of the topic distribution. The impact was influenced by the set of hyper-parameters, the average text length, and in some situations, on the number of documents. In general, parameter estimates for both were more accurate when hyperparameters were $\alpha = 0.5$ & $\beta = 0.05$. Although the influence of hyper-parameters tends to decrease as the text length increased.

Additionally, results suggested that Gibbs sampling algorithm was more accurate for estimating topic proportions than VEM when the average text length was small. Accuracy differences between parameter estimate, however, decreased as the average text length increased. These results suggest the VEM algorithm does about

as well the Gibbs sampling while using less time and less computational resources. Furthermore, contrary to previous evidence (e.g., Syed & Spruit, 2018; Tang et al., 2014), this simulation study showed that LDA results can be accurate for corpora with small average text length.

This study provides useful information about the performance of two estimators for LDA models under conditions typical of CR answers on test items. Future studies would be useful to investigate the effect of the Dirichlet priors for these estimators. Future research also could compare the results of this study using real data.

References

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518), 859–877.
- Cardozo-Gaibisso, L., Kim, S., Buxton, C., & Cohen, A. (2019). Thinking beyond the score: Multidimensional analysis of student performance to inform the next generation of science assessments. *Journal of Research in Science Teaching*, 57, 856–878.
- Chen, Q., Yao, L., & Yang, J. (2016). Short text classification based on LDA topic model. In *Proceedings of the 2016 International Conference on Audio, Language, and Image Processing* (pp. 749–753).
- Choi, H. -J., Kwak, M., Kim, S., Xiong, J., Cohen, A. S., & Bottge, B. A. (2019). An application of a topic model to two educational assessments. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative Psychology: The 83rd Annual Meeting of the Psychometric Society* (pp. 449–459).
- Cohen, A. S., Kane, M. T., & Kim, S. H. (2001). The precision of simulation study results. *Applied Psychological Measurement*, 25(2), 136–145.
- Crossley, S., Kyle, K., Davenport, J., & McNamara, D. S. (2016). *Automatic assessment of constructed response data in a chemistry tutor*. International Educational Data Mining Society.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407.
- Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101, 5228–5235.
- Grün, B., & Hornik, K. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(1), 1–30.
- Hellman, S., Murray, W. R., Wiemerslage, A., Rosenstein, M., Foltz, P., Becker, L., & Derr, M. (2020). Multiple instances learning for content feedback localization without annotation. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 30–40).
- Hu, X., Sun, N., Zhang, C., & Chua, T. S. (2009, November). Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (pp. 919–928).
- Kim, J. N. (2020). Variational expectation-maximization algorithm in posterior distribution of a latent Dirichlet allocation model for research topic analysis. *Journal of Korea Multimedia Society*, 23(7), 883–890.
- Kim, S., Kwak, M., Cardozo-Gaibisso, L., Buxton, C., & Cohen, A. S. (2017). Statistical and qualitative analyses of students' answers to a constructed response test of science inquiry knowledge. *Journal of Writing Analytics*, 1, 82–102.

- Kwak, M. (2019). Parameter recovery in latent Dirichlet allocation (LDA): Potential utility of LDA in formative constructed response assessment. Unpublished doctoral dissertation.
- Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement, 33*, 353–373.
- Lissitz, R. W., Hou, X., & Slater, S. C. (2012). The contribution of constructed response items to large scale assessment: Measuring and understanding their impact. *Journal of Applied Testing Technology, 13*(3).
- Ponweiser, M. (2012). Latent Dirichlet allocation in R (Doctoral dissertation). Retrieved from <https://epub.wu.ac.at/3558/1/main.pdf>
- Shin, J., Guo, Q., & Gierl, M. J. (2019). Multiple-choice item distractor development using topic modeling approaches. *Frontiers in Psychology, 10*, 1–14.
- Syed, S., & Spruit, M. (2018). Selecting priors for latent Dirichlet allocation. In *2018 IEEE 12th International Conference on Semantic Computing* (pp. 194–202).
- Taddy, M. (2012). On estimation and selection for topic models. In *Artificial intelligence and statistics* (pp. 1184–1193).
- Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014) Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of 2014 International Conference on Machine Learning* (pp. 190–198).
- Tietjen, G. L. (2012). *A topical dictionary of statistics*. Springer.
- Wallach, H. M., Mimno, D. M., & McCallum, A. (2009). Rethinking LDA: Why priors matter. In *Advances in neural information processing systems* (pp. 1973–1981).
- Walther, B. A., & Moore, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography, 28*(6), 815–829.
- Wheeler, J. M., Cohen, A. S., Xiong, J., Lee, J., & Choi, H. J. (2021). Sample size for latent Dirichlet allocation of constructed-response items. In *Quantitative psychology* (pp. 263–273). Springer.
- Xiong, J., Choi, H.-J., Kim, S., Kwak, M., & Cohen, A. S. (2019). Topic modeling of constructed-response answers on social study assessments. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology: The 84th annual meeting of the psychometric society* (pp. 263–274). Springer.
- Zuo, Y., Zhao, J., & Xu, K. (2016). Word network topic model: A simple but general solution for short and imbalanced texts. *Knowledge and Information Systems, 48*, 379–398.

Relationship Between Students' Test Results and Their Performance in Higher Education Using Different Test Scores



Marie Wiberg, Juan Li, Per-Erik Lyrén, and James O. Ramsay

Abstract The aim of this study is to examine the relationship between students' college admissions test results and their performance in higher education using sum scores and optimal full-data scores. We used students from four university programs to examine predictive validity in terms of how the students performed on their studies in terms of obtained credits, as compared with their college admissions test results. The students' test results were calculated using the commonly used sum scores and the recently proposed optimal scores. We also examined the predictive validity of the test scores while controlling for the student backgrounds in terms of educational background, migration background, and gender. The results show that using optimal scores or sum scores yields slightly different test score distributions, especially the score distribution among the highest test performers differed. Practical implications of which test scores to use in college admissions testing in the future are discussed.

Keywords Optimal scores · Sum scores · Predictive validity · College admissions test

M. Wiberg (✉)

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

J. Li

Ottawa Hospital Research Institute, Ottawa, ON, Canada

P.-E. Lyrén

Department of Applied Educational Science, Umeå University, Umeå, Sweden

J. O. Ramsay

Department of Psychology, McGill University, Montreal, QC, Canada

1 Introduction

Sum scores are used in most standardized tests across the world, for example the Graduate Record Exam (GRE) (GRE, 2021), the Scholastic Aptitude Test (SAT) produced by the College Board, the aptitude tests produced by American College Testing (ACT) (Dorans, 1999), and the Swedish scholastic aptitude test (SweSAT) (Lyrén et al., 2014). Recently, optimal scoring was proposed as an alternative to sum scores which takes care of the different information provided by different items (Ramsay & Wiberg, 2017a,b). An advantage with optimal scores is that the scores becomes fairer in comparison to the actual knowledge level and thus high achievers may achieve higher optimal scores than sum scores as seen in Wiberg et al. (2019).

In Sweden, the selection component in admissions to university (the other component being eligibility) is based on a quota system. A certain proportion of candidates are admitted from quota groups based on different selection instruments, where grades from upper-secondary school (USGPA) and scores from the optional admissions test, SweSAT, are the two most common ones. Candidates who have both a valid USGPA (which most candidates have) and a valid SweSAT score will be placed in both quota groups, so taking the SweSAT can only increase one's chances of being admitted. If a test taker has several valid SweSAT scores, the best score is used in the admissions.

The goal of the SweSAT is to select those students who are most likely to perform well in higher education. Consequently, as is the case with any other selection instrument, the predictive validity of the scores is central to the overall validity of the use and interpretation of SweSAT test scores. Predictive validity studies on selection instruments in Sweden have often compared the predictive strength of the USGPA and SweSAT scores. The most common finding is that the USGPA is a better predictor than SweSAT scores (Svensson et al., 2001; Cliffordson, 2008; Cliffordson & Askling, 2006; Lyrén, 2008) and that the predictive strength differs between university programs for both instruments. For example, Lyrén et al. (2014) analyzed eleven different programs and found that the correlation between SweSAT scores and the performance criterion was non-significant for two programs (medicine and social work) and that it varied between 0.2 and 0.4 for the other nine programs (engineering, nursing, economics, teaching, etc.). They also found that the correlations were similar for the two section scores (Verbal and Quantitative), except for engineering programs where the correlations were higher for the Quantitative score than for the Verbal score.

In this paper we are interested in examining the predictive validity of the SweSAT if we use full information optimal scoring (Ramsay et al., 2020) as compared with using sum scores. Previous studies with optimal scores have focused on examining the possibility to use optimal scores instead of sum scores when we have binary scored multiple choice items (Ramsay & Wiberg, 2017b,a) and also a comparison between full information optimal score and binary sum scores (Wiberg et al., 2018) as well as a comparison between binary optimal scores and item response theory scores (Wiberg et al., 2019). In this paper we use full information optimal scores as

described in Ramsay et al. (2020). It is called full information optimal score because information in both correct and incorrect responses was used for scoring. This paper is different from previous papers as the focus is not to refine optimal scoring but instead to examine the predictive validity of the optimal scores in terms of how students perform once they have been admitted to a university program of their choice. The overall aim is to examine the predictive validity and thus the relationship between students' college admissions test results and the students' performance in higher education using sum scores and optimal scores.

The rest of this paper is structured as follows: Next, the method section with the different test scores, the sample used and the statistical analysis are described. This section is followed by a result section and the paper ends with a discussion with some concluding remarks.

2 Method

2.1 Test Scores

We focus on two different kinds of test scores; sum scores and full information optimal scores. For multiple choice items, sum scores are typically defined as the number of items the test taker answered correctly. The full information optimal score, further referred to as optimal scores, is formally defined in Ramsay et al. (2020) and thus only briefly described here. In the later empirical study, we use the freely available software R and especially the package TestGardener (Ramsay & Li, 2021) to estimate the optimal scores. The initial proposal of optimal scores were made by Ramsay and Wiberg (2017b).

The basic idea is to estimate the scores based on the interaction between the performance of items/options and test taker; and using surprisal $W_{im}(\theta) = -\log_M P_{im}(\theta)$ rather than probability $P_{im}(\theta)$ in the estimation process, where θ is the given test taker's ability and M is the number of options of item i . Let $m = 1, \dots, M_i$ represent the different answer options for item i , and let $P_{im}(\theta)$ be the probability of a test taker with ability θ choosing the m option. The multinomial item response function can then be defined as

$$P_{im}(\theta) = \frac{\exp [W_{im}(\theta)]}{\sum_{l=1}^{M_i} \exp [W_{il}(\theta)]} \quad (1)$$

where W_{im} is an unbounded function associated with the m th answer option for item i . We see in this formulation two actions: (1) the exponential transform that ensures that the probability will be positive, and (2) the normalization by dividing by sum $\exp(W)$ in order to ensure that the probability values sum to one. The optimal score is found by minimizing the value of θ as defined by the following equation

$$\frac{dH}{d\theta} = - \sum_{i=1}^n \left[\sum_{m=1}^{M_i} [U_{im} - P_{im}(\theta)] \frac{dW_{im}}{d\theta} \right] = 0, \quad (2)$$

where $U_{im} - P_{im}$ is the difference between the data and the model fit and

$$dW_{im}/d\theta$$

is a coefficient that gives more weight if the item option contributes more to the knowledge of the test taker's ability. For more computational details of full information optimal scores, please refer to Ramsay et al. (2020).

2.2 The SweSAT

To examine student performance and to predict their success in university we used scores from the college admissions test SweSAT. The test is optional and is given twice a year. The test taker can repeat the test as many times as they prefer as only the best results counts. The test results are valid for five years. The SweSAT contains 160 multiple choice items and is divided into one verbal section and one quantitative section with 80 items each. The verbal section contains Vocabulary (20 items), Swedish reading comprehension (20 items), English reading comprehension (20 items) and Sentence completion (20 items). The quantitative section contained Data sufficiency (12 items), Diagrams, tables and maps (24 items), Mathematical problem solving (24 items) and Quantitative comparisons (20 items). Sum scores are used to calculate the test takers score on the test. There are 4–5 response alternatives to each of the multiple-choice items.

2.3 Participants

We used samples of students who were admitted to four different higher education programs in Sweden. The programs were chosen as they have different variations in their test score distributions and the chosen programs were: biomedical analysts (biomed), civil engineering (cing), college engineering (hing), and medical program (medical). The distribution of students at the different examined programs are given in the result section in Table 1.

The following student background variables were examined. Migration defined as 1 if the student or at least one of the students' parents were born in Sweden and 0 otherwise. Boys were coded as 1 and girls were coded as 0. Educational background was coded as 0 if the student had high school education or lower, and it was coded as 1 if the student had any post high school education. As it is also possible to get admitted to a university program in Sweden using only high school

grades, we used a grade variable which was composed by the final high school grade for each student. The grade variable was a constructed grade variable from the years 1997–2012, as we had high school graduates from all those years. The constructed grade variable was used so that all students were placed on the same grading scale even though the grade scale has changed in Sweden during those years. A grade A is equivalent to 20, a grade B is equivalent to 17.5, a grade C is equivalent to 15, a grade D is equivalent to 12.5, grade E is equivalent to 10 and the grade F is equivalent to 0. A student can also get extra credits (0.40) for extra curriculum activities. This means that the grade point average has a range of 0.0–20.40.

To get a measure of the students' achievement on their college education program we used a constructed variable, *Relprest* which have been used in other validation studies (e.g. Lyrén et al., 2014). *Relprest* is defined as the ratio between the students' passed credits and registered credits in their first year of college or university. The range of *Relprest* is 0.0–2.0, as students get zero if they do not take any of the credits they signed up for and some students have signed up for twice as many credits as the normal study rate.

2.4 Statistical Analysis

In the analyses we used both SweSAT sum scores and SweSAT optimal scores. We started by examining the score distributions using histograms and to examine the linear relationship between the two test scores we used scatterplots. Next, we examined the linear relationship between the test scores and *Relprest* with Pearson correlation. To examine the possible predictive effect, we used linear regressions with *Relprest* as dependent variable and the different test scores together with the students' background variables as independent variables. We also examined the test score distributions of the top 10% students with respect to their sum scores. The optimal scores were calculated using TestGardener (Ramsay & Li, 2021; Li et al., 2019) and the other statistical analyses were done in SPSS.

3 Results

Figure 1 displays the test score distributions for sum scores and optimal scores and Fig. 2 gives the scatterplot for the whole sample of those who took the SweSAT. From this figure it is clear that the distributions are not exactly the same. The distributions however share some similar features as the mean of the sum scores was 94.26 (SD = 21.98, Range: 32–151) and the mean for optimal scores was 94.39 (SD = 21.95, Range: 40.12–142.79). Although the sum scores have lower minimum and higher maximum than the optimal scores, the mid score range is a bit more flatten for the optimal score distribution as compared with the sum score distribution. The upper score range also differed depending on used test score.

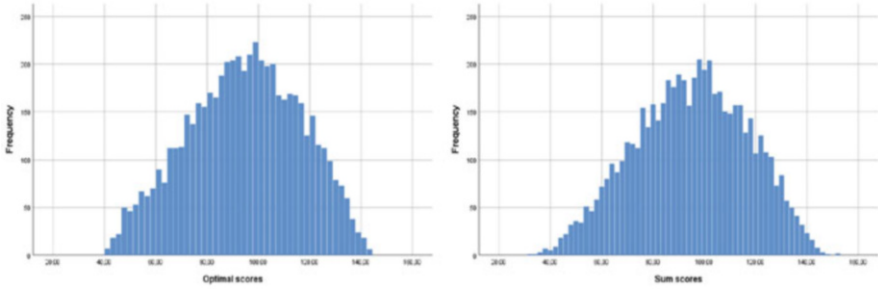


Fig. 1 Test score distributions with optimal scores to the left and sum scores to the right

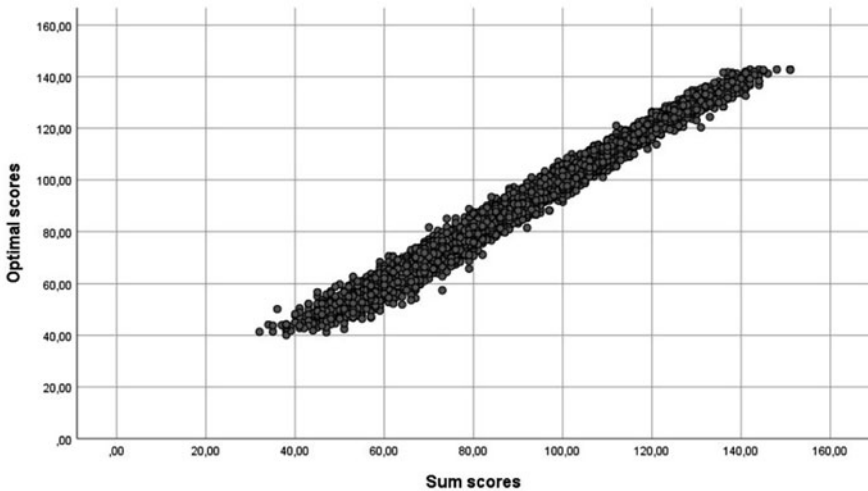


Fig. 2 Scatterplot between optimal scores and sum scores

The left part of Table 1 shows the correlation between full information optimal scores, sum scores and *Relprest* for the total SweSAT and the two SweSAT subsections; Quantitative and Verbal. From this table it is evident that there are overall quite small differences between optimal scores and sum scores. The non-significant correlations for the medical program are probably due to the fact that the variations of test scores are small in the medical programs. The right part of Table 1 gives the correlation between the students' high school grades and *Relprest*. Again, the weak correlation for medical students is due to the small variation of high school grades in this group.

To study the linear correlation between the SweSAT optimal score and the SweSAT sum scores we refer to Fig. 3 for the four university programs of interest. The correlations were very high and ranged between 0.98 (Biomed and Hing) to 0.99 (Cing and Medical). From these numbers and Fig. 3, it is evident that the sum scores and optimal scores are highly correlated but the scores differ for most of the

Table 1 Correlations between full information optimal scores, sum scores and *Relprest* in the left columns for the total SweSAT and the two subsections. The right columns shows the correlations between grades and *Relprest*

Exam	n	Total		Quant		Verbal		Grade		
		r	Sig.	r	Sig.	r	Sig.	n	r	Sig.
<i>Biomed</i>										
Optimal	178	0.39	***	0.30	***	0.38	***	149	0.44	***
Sum	178	0.38	***	0.28	***	0.38	***			***
<i>Cing</i>										
Optimal	3172	0.22	***	0.26	***	0.12	***	3036	0.38	***
Sum	3172	0.21	***	0.25	***	0.12	***			***
<i>Hing</i>										
Optimal	1404	0.20	***	0.22	***	0.13	***	1297	0.38	***
Sum	1404	0.19	***	0.21	***	0.13	***			***
<i>Medical</i>										
Optimal	827	-0.03	NS	0.00	NS	-0.05	NS	740	0.24	***
Sum	827	0.01	NS	0.01	NS	-0.05	NS			***

Biomed = Biomedical analytics, Cing = civil engineering, Hing = College engineering, Medical = Medical program. Total = Total SweSAT scores. Quant = Quantitative section scores, Verb = Verbal section scores. Grade = Correlation between *Relprest* and grades. NS = non-significant
 *** = *p*-value less than 0.01

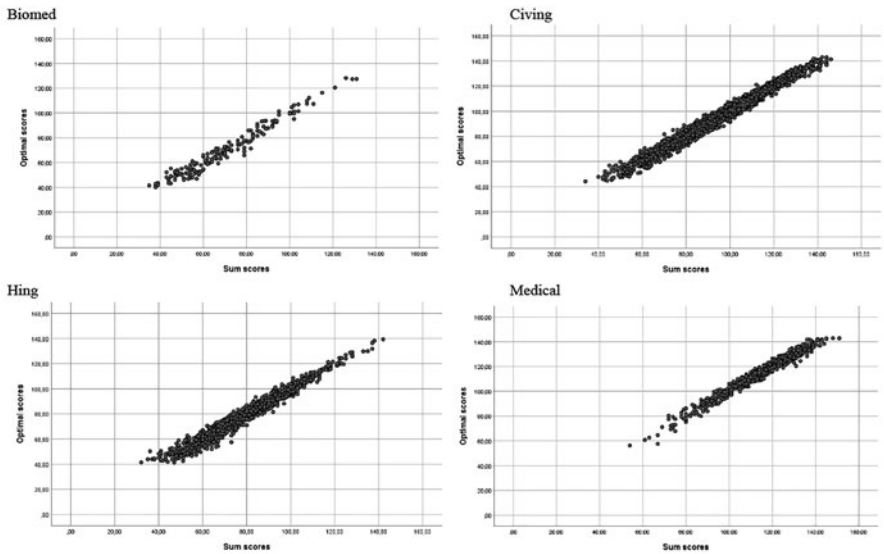


Fig. 3 The relationship between SweSAT optimal scores and SweSAT sum scores in the four different university programs

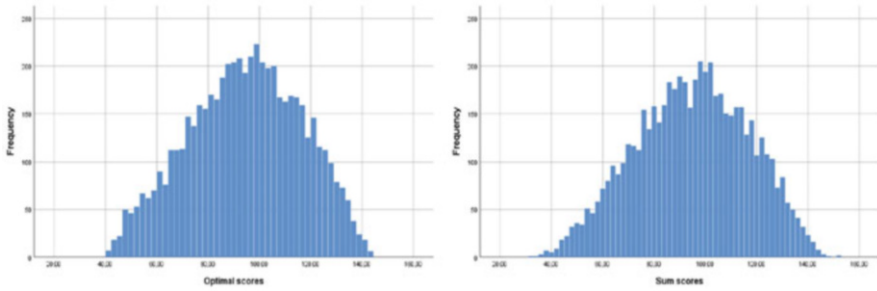


Fig. 4 Optimal test scores and sum score distributions for top 10% performers chosen from SweSAT sum scores

test takers. The differences differ over the score range depending on the university program and the score difference can be as large as 10 score points.

As SweSAT is primarily used as a higher education admissions test, higher scores are of more interest than the lower score range. The 10% top performers admitted to the programs as defined from the sum scores are given in Fig. 4. From these plots it is evident that the top performers have slightly different test score distributions.

To further examine the predictive validity of the SweSAT scores we used the variable *Relprest* which is an indication of how the admitted students performed in their first year of university in comparison to what courses they have signed up for. We examined several student background variables, but the students' gender or educational background was never significant in any of the examined programs and thus the result is excluded from the table. The reason for the non-significance of the educational background is probably due to the rough definition of this variable as it only stated whether or not you have studied anything after high school or not. In Table 2, linear regressions with *Relprest* as dependent variable with optimal scores on every second line and sum scores on the other lines are given. We examined three different linear regression models. In model 1, only either optimal scores or sum scores were used as independent variable. In model 2, we used the test scores together with the grade variable. Finally, in model 3 we included the test scores, grades and the students' migration home background. The best fitting model was model 2 for all university programs and for the total SweSAT as well as for the subsections. Model 3 had many non-significant variables for most of the examined programs, regardless of test score used. We give the value of R from the linear regressions in the table and note that the values are very similar regardless of the test score used. Only small differences are shown and those were mainly when model 1 was used. Again, a reason for the non-significant values for the medical program is probably due to the very small variation in test scores for those who got admitted to the program.

Table 2 The R values and the sample sizes (n) from the three different linear regression models with *Relprest* as dependent variable with optimal scores on every second line and sum scores the other lines

Exam	n	Model 1			Model 2 (G)			Model 3 (GI)		
		Tot	Q	V	Tot	Q	V	Tot	Q	V
<i>Biomed</i>										
Optimal	178	0.39**	0.30**	0.38**	0.46**	0.44	0.47**	0.47	0.47	0.49
Sum	178	0.38**	0.28**	0.38**	0.46**	0.44	0.48**	0.47	0.47	0.49
<i>Cing</i>										
Optimal	3172	0.22**	0.26**	0.12**	0.39**	0.41**	0.39	0.39	0.41*	0.39
Sum	3172	0.21**	0.25**	0.12**	0.39**	0.41**	0.38	0.39	0.41*	0.39
<i>Hing</i>										
Optimal	1404	0.20**	0.22**	0.13**	0.39**	0.40**	0.38**	0.40**	0.41**	0.39
Sum	1404	0.20**	0.21**	0.13**	0.39**	0.40**	0.38**	0.40**	0.41**	0.39
<i>Medical</i>										
Optimal	827	0.03	0.00	0.05	0.24	0.24	0.24	0.25	0.25	0.25
Sum	827	0.02	0.01	0.05	0.24	0.24	0.24	0.25	0.25	0.25
<i>All</i>										
Optimal	5581	0.25**	0.26**	0.19**	0.40**	0.40**	0.40**	0.40**	0.41**	0.40**
Sum	5581	0.25**	0.26**	0.19**	0.40**	0.40**	0.39**	0.40**	0.41**	0.40**

Biomed = Biomedical analytics, Cing = civil engineering, Hing = College engineering, Medical = Medical program. N = Number of test takers. Tot = Total test score. Q = Quantitative test score, V = Verbal test score. Sum = Sum scores are used as independent variable instead of optimal scores. G = Grade, I = Immigration

** = *p*-value less than 0.01, * = *p*-value less than 0.1

4 Discussion

The overall aim was to examine the predictive validity and thus the relationship between students’ college admissions test results and the students’ performance in higher education using sum scores and optimal scores. The results indicated that both optimal scores and sum scores can predict the students’ university performance similarly regardless if we control for some covariates or not. Although the test score distributions differed, the overall results of predictivity of the students’ performance were similar. This is good news as it means that optimal scores can be used in these situations. Although the overall conclusions were similar, the test score distributions differed in the sense that the optimal score distribution had a slightly more flattened curve than the sum score distribution. This means that for a certain student it may have impact which test score is used when the test score is used for selection to higher education even though a clear difference is not seen on the overall results. The differences between test results based on sum scores and optimal scores are typically larger for programs which require high test scores. The result that different test takers may get different sum scores and optimal scores are inline with previous

studies of optimal scores (Ramsay & Wiberg, 2017a,b; Wiberg et al., 2018). The impact for a specific student should be addressed further in the future.

There were a few limitations of this study. First, as a measure of success in higher education we used a relative performance measure of the students' performance. This measure is probably a bit blunt and thus future studies should probably use a more refined measurement. However, this measure was used in the SweSAT prognosis study by Lyrén et al. (2014), which obtained similar result for the sum scores as in our study. Note, some numbers concerning the sum scores differed between our study and their study as we in contrast to them only used complete cases. Second, we only had access to a few student background variables and in future studies, it would be of interest to include other background variables. Third, in this study we only had access to those admitted to the university programs and thus the study has a range restriction. As the optimal scores and sum scores differ in their distributions, it is likely that the rank of the students differ within the test scores. If one would change from sum scores to optimal scores it is likely that some students may have not admitted to a program and others were admitted to a program and thus the choice of test score could potential influence the life of a student. However, on a group level the results are similar and thus one should be comfortable to use either sum scores or optimal scores in admissions tests. An advantage of using optimal scores in sum scores, seen in e.g. Ramsay and Wiberg (2017b), is that the precision of estimating the ability of the students is better and thus optimal scores should be considered for high stakes test as it would be a fairer instrument for the students.

Acknowledgement The research was funded by the Swedish Wallenberg MMW 2019.0129 grant.

References

- Cliffordson, C. (2008). Differential prediction of study success across academic programs in the swedish context: The validity of grades and tests as selection instruments for higher education. *Educational Assessment, 13*(1), 56–75.
- Cliffordson, C., & Askling, B. (2006). Different grounds for admission: Its effects on recruitment and achievement in medical education. *Scandinavian Journal of Educational Research, 50*(1), 45–62.
- Dorans, N. J. (1999). Correspondences between act and sat®i scores. *ETS Research Report Series, 1999*(1), i–18.
- GRE (2021). *Guide to the Graduate Record Exams (GRE)*. https://www.ets.org/s/gre/pdf/gre_guide.pdf. Accessed 19 Jan 2021.
- Li, J., Ramsay, J. O., & Wiberg, M. (2019). TestGardener: A program for optimal scoring and graphical analysis. In *Quantitative Psychology 83rd Annual Meeting of the Psychometric Society* (pp. 87–94). Springer.
- Lyrén, P.-E. (2008). Prediction of academic performance by means of the swedish scholastic assessment test. *Scandinavian Journal of Educational Research, 52*(6), 565–581.
- Lyrén, P.-E., Rolfzman, E., Wedman, J., Wikström, C., & Wikström, M. (2014). *Det nya högskoleprovet: samband mellan provresultat och prestation i högskolan [The new SweSAT: Association between test results and performance in higher education]*. Umeå, Sweden: Department of Applied Educational Science, Umeå University.

- Ramsay, J. O., & Li, J. (2021). TestGardener: Optimal Analysis of Test and Rating Scale Data. R package version 2.0.1 <https://CRAN.R-project.org/package=TestGardener>
- Ramsay, J. O., & Wiberg, M. (2017a). Breaking through the sum scoring barrier. In *Quantitative Psychology 81st Annual Meeting of the Psychometric Society* (pp. 151–158). Springer.
- Ramsay, J. O., & Wiberg, M. (2017b). A strategy for replacing sum scoring. *Journal of Educational and Behavioral Statistics*, 42(3), 282–307.
- Ramsay, J. O., Wiberg, M., & Li, J. (2020). Full information optimal scoring. *Journal of Educational and Behavioral Statistics*, 45(3), 297–315.
- Svensson, A., Gustafsson, J.-E. and Reuterberg, S.-E. (2001). *Högskoleprovets prognosvärde. Samband mellan provresultat och framgång första året vid civilingenjörs-, jurist- och grundskollärautbildningarna* [The prognostic value of the SweSAT. Association between test result and first-year academic performance at the civil engineering, law and elementary school teacher programmes] (Högskoleverkets rapportserie No. 2001:19 R). Stockholm: Högskoleverket.
- Wiberg, M., Ramsay, J. O., & Li, J. (2018). Optimal scores as an alternative to sum scores. In *Quantitative Psychology 82nd Annual Meeting of the Psychometric Society* (pp. 1–10). Springer.
- Wiberg, M., Ramsay, J. O., & Li, J. (2019). Optimal scores: An alternative to parametric item response theory and sum scores. *Psychometrika*, 84(1), 310–322.

Statistical Properties of Lower Bounds and Factor Analysis Methods for Reliability Estimation



Julius M. Pfadt  and Klaas Sijtsma 

Abstract In this study, we compared the numerical performance of reliability coefficients based on classical test theory and factor analysis. We investigated the coefficients' divergence from reliability and their population values using unidimensional and multidimensional data generated from both an item response theory and a factor model. In addition, we studied reliability coefficients' performance when the tested model was misspecified. For unidimensionality, coefficients α , λ_2 , and coefficient ω_u approximated reliability well and were almost unbiased regardless of the data-generating model. For multidimensionality, coefficient ω_t performed best with both data generating models. When the tested model was unidimensional but the data multidimensional, all coefficients underestimated reliability. When the tested model incorrectly assumed a common factor in addition to group factors but the data was purely multidimensional, coefficients ω_h and ω_t identified the underlying data structure well. In practice, we recommend researchers use reliability coefficients that are based on factor analysis when data are multidimensional; when data are unidimensional both classical test theory methods and factor analysis methods get the job done.

Keywords Classical test theory · Coefficient alpha · Coefficient lambda2 · Coefficient lambda4 · Coefficient omegaH · Coefficient omegaT · Coefficient omegaU · Factor analysis · Greatest lower bound

J. M. Pfadt (✉)

Ulm University, Department of Psychological Research Methods, Ulm University, Ulm, Germany
e-mail: julius.pfadt@uni-ulm.de

K. Sijtsma

Department of Methodology and Statistics TSB, Tilburg University, Tilburg, The Netherlands
e-mail: k.sijtsma@tilburguniversity.edu

1 Introduction

Most methods for reliability estimation fall into methods based on classical test theory (CTT; Lord & Novick, 1968) and factor analysis (FA; e.g., Jöreskog, 1971). We studied sampling properties of lower bound coefficients α , λ_2 , λ_4 , and the greatest lower bound (GLB) from CTT, and coefficients ω_u , ω_h , and ω_t from FA. We ran a simulation study assessing statistical properties of these estimators, using item response theory (IRT) and FA models to generate unidimensional and multidimensional data while controlling the reliability. As far as we know, this comparison is novel. We investigated two special cases in which the population model is a multidimensional FA-model, but the tested model is misspecified. Before we discuss the process of data generation, we describe the reliability coefficients α , λ_2 , λ_4 , *GLB*, ω_u , ω_h , and ω_t .

1.1 CTT Coefficients

Based on the CTT layout of the test score, $X = T + E$ (X : test score, sum of item scores; T : true score; E : random measurement error; see Sijtsma & Pfadt, 2021), we studied coefficients α , λ_2 , λ_4 , and the *GLB*. These coefficients approximate reliability, ρ , defined as the proportion of test-score variance that is true-score variance in a population; that is, $\rho = \sigma_T^2 / \sigma_X^2$. The approximations are theoretical lower bounds to the reliability. The coefficients' equations are the following (for further information, see, e.g., Sijtsma & Van der Ark, 2021). For coefficient α , let σ_X^2 be the test score variance and let σ_j^2 be the variance of item j ($j = 1, \dots, J$); then

$$\alpha = \frac{J}{J-1} \left(1 - \frac{\sum_{j=1}^J \sigma_j^2}{\sigma_X^2} \right). \quad (1)$$

For coefficient λ_2 , let σ_{jk}^2 be the covariance between items j and k , then

$$\lambda_2 = 1 - \frac{\sum_{j=1}^J \sigma_j^2 - \sqrt{\frac{J}{J-1} \sum \sum_{j \neq k} \sigma_{jk}^2}}{\sigma_X^2}. \quad (2)$$

For coefficient λ_4 , split a test in two item subsets without overlap and not necessarily equally sized, call this split partition P , and let σ_A^2 and σ_B^2 be the variances of the test scores on each subset. Then, coefficient α for this partition, $\alpha(P)$, equals: $\alpha(P) = 2 \cdot \left(1 - \frac{\sigma_A^2 + \sigma_B^2}{\sigma_X^2} \right)$. Coefficient λ_4 is the greatest value of $\alpha(P)$ across all partitions P ; that is,

$$\lambda_4 = \max_P [\alpha(P)]. \quad (3)$$

Let Σ_X be the covariance matrix that is split into $\Sigma_X = \Sigma_T + \Sigma_E$, where Σ_T contains the true score variances and diagonal Σ_E contains the error score variances. The estimates for Σ_T and Σ_E are found in an iterative procedure where the trace of Σ_E is maximized while the matrices Σ_T and Σ_E stay positive semi-definite (Woodhouse & Jackson, 1977). Then

$$GLB = 1 - \frac{tr(\Sigma_E)}{\sigma_X^2}. \tag{4}$$

1.2 FA Coefficients

The FA approach to reliability is based on the assumption that the CTT model, $X = T + E$, can be substituted with the FA-model. Note that the models are different and thus define different reliability conceptions except for extreme cases. For items, the CTT model is $X_j = T_j + E_j$, with $T_j = \mathcal{E}(X_{jr})$, r indexes independent replications of item j , comparable with the stochastic subject formulation of response probability in IRT (Holland, 1990). FA models item scores as $X_j = b_j + \sum_{q=1}^Q a_{jq}\xi_q + \delta_j$, with intercept b_j , latent variable ξ_q indexed q ($q = 1, \dots, Q$), a_{jq} the loading of item j on latent variable ξ_q , and δ_j the residual consisting of unexplained item components and random measurement error. One may notice that different items have one or more latent variables in common, whereas in CTT, true scores unique to the item lack additional modeling. FA extracts one or more factors from the data. These factors predict the outcome variables, here, the items. The level of prediction is represented by loadings that link each item with one or more factors.

The FA approach to reliability is based on the assumption that the sum of the squared factor loadings approximates the true-score variance of items. The residual variances, the part of items the factor(s) cannot predict, substitute the error-score variance. Reliability is defined as the proportion of the test-score variance that is due to one or more common factors. For instance, the single-factor model (Spearman, 1904) describes the data matrix \mathbf{X} of multivariate observations as

$$\mathbf{X} = \mathbf{g}\mathbf{c}^T + \mathbf{E}, \tag{5}$$

where \mathbf{c} denotes the factor loadings on one common factor \mathbf{g} (replacing general notation ξ) and \mathbf{E} the matrix of residuals, the part of the item scores that the common factor cannot explain. Since the residuals are assumed independent, the covariance matrix of \mathbf{E} is diagonal and has elements e_j representing residual variances. Coefficient ω_u (u for unidimensional; McDonald, 1999) equals

$$\omega_u = \frac{(\sum c)^2}{(\sum c)^2 + \sum e}. \tag{6}$$

When data are multidimensional, one can either estimate reliability for each subscale using ω_u or one can employ coefficients ω_h (h for hierarchical) and ω_t (t for total; Zinbarg et al., 2005). Therefore, consider the following multidimensional bi-factor model,

$$\mathbf{X} = \mathbf{g}\mathbf{c}^T + \mathbf{F}\mathbf{A}^T + \mathbf{E}, \quad (7)$$

where \mathbf{A} denotes the $J \times Q$ loading matrix for the Q group factors collected in \mathbf{F} . The Q group factors are common to some items but not all. The residual variances e_j of the residual matrix \mathbf{E} represent the part of the items that the common factor and the group factors cannot explain. Coefficient ω_h equals

$$\omega_h = \frac{\mathbf{c}^T \mathbf{c}}{\mathbf{c}^T \mathbf{c} + \mathbf{1}_Q^T \mathbf{A}^T \mathbf{A} \mathbf{1}_Q + \sum \mathbf{e}}, \quad (8)$$

where $\mathbf{1}_Q$ is a $Q \times 1$ sum vector. The coefficient describes the common factor saturation of a test in the presence of group factors. The value of coefficient ω_h addresses the question “how well does a multidimensional scale represent a common attribute”. Coefficient ω_h is not an estimate of the reliability as indicated by coefficient ω_t , because coefficient ω_h does not relate all true score variance to the total variance of a test, but only the variance due to a general factor. Coefficient ω_t equals

$$\omega_t = \frac{\mathbf{c}^T \mathbf{c} + \mathbf{1}_Q^T \mathbf{A}^T \mathbf{A} \mathbf{1}_Q}{\mathbf{c}^T \mathbf{c} + \mathbf{1}_Q^T \mathbf{A}^T \mathbf{A} \mathbf{1}_Q + \sum \mathbf{e}}, \quad (9)$$

and describes the proportion of variance in the test that is due to a common attribute and specific attributes that pertain to subsets of items, which is the true-score variance. The loadings and residual variances for the ω -coefficients can be obtained from both a confirmatory factor model and an exploratory factor model.

2 Simulation Study

We compared reliability coefficients estimated in samples of simulated data with the reliability of the population model that generated the data. Oosterwijk et al. (2017) compared several lower bounds with population reliability by generating data from a two-dimensional graded response model (GRM; Samejima, 1968). From the GRM parameters, they computed the item true scores and then the reliability. Zinbarg et al. (2006) used a factor model to generate data for the evaluation of estimation methods for coefficient ω_h . Assuming the factor variance represents the true score variance, one can obtain the population reliability from the factor model parameters.

To rule out the possibility that the data generation process confounds the outcomes, we generated data both with the GRM and a factor model, and evaluated the lower bound coefficients α , λ_2 , λ_4 , and the *GLB*, and the FA-coefficients ω_u , ω_h , and ω_r . Furthermore, we investigated the ramifications of a misspecified model when estimating reliability coefficients. We generated data based on researchers (1) overlooking a scale’s multidimensionality, and (2) incorrectly assuming presence of a common attribute.

2.1 Method

Data Generation from GRM

The data generation based on the GRM in this study is similar to the data generation in Oosterwijk et al. (2017). The GRM defines for each polytomous item j a slope parameter a_j and a location parameter b_{jx} for each item score x . The cumulative response probability of scoring at least x ($x = 0, \dots, m$) on item j as a function of latent variable(s) θ_q ($q = 1, \dots, Q$) collected in θ is expressed as

$$P(X_j \geq x | \theta) = \frac{\exp \left[\sum_{q=1}^Q a_{jq} (\theta_q - b_{jx}) \right]}{1 + \exp \left[\sum_{q=1}^Q a_{jq} (\theta_q - b_{jx}) \right]}. \tag{10}$$

The response probability of scoring exactly x on item j is given by

$$P(X_j = x | \theta) = P(X_j \geq x | \theta) - P(X_j \geq x + 1 | \theta), \tag{11}$$

with $P(X_j \geq 0 | \theta) = 1$ and $P(X_j > m | \theta) = 0$ for response categories $0, \dots, 4$. For our study, we chose $Q = 1$ and $Q = 3$. The model definition and data generation for the GRM largely follows Oosterwijk et al. (2017) but for clarity we reiterate most of it here.

Unidimensional Model To model a wide range of person parameters, we defined $\theta = -5, -4.95, -4.9, \dots, 4.9, 4.95, 5$, that is, 201 evenly spaced values in total that followed a standard normal distribution. Numbers of items were $J = 9, 18$, item scores were $x = 0, \dots, 4$, and slope parameters $a_j \in U(1, 1.5)$. Item location parameters were $b_{jx} = \tau_j + \kappa_x$, with $\tau_j = (j-1)/(J-1) - .5$ and $\kappa_x = (-1.5, -0.5, 0.5, 1.5)^T$, with $x = 0, \dots, 4$ for five item scores.

Multidimensional Model We chose $Q = 3$, identical equally-spaced person parameter vectors $\theta_1, \theta_2, \theta_3$ as in the unidimensional case, and a multivariate normal distribution with means 0, variances 1, and correlations .3. Numbers of items and item scores were the same as for the unidimensional model. Multidimensionality was achieved by assigning slopes $a_{jq} \in U(1, 2)$ for latent variable θ_q to a third of the items and $a_{jq} = 0$ to the other items, and so on. For

example, for $J = 9$, $\mathbf{a}_1 = (1.75, 1.25, 1.55, \mathbf{0})^T$, $\mathbf{a}_2 = (\mathbf{0}, 1.60, 1.04, 1.30, \mathbf{0})^T$, and $\mathbf{a}_3 = (\mathbf{0}, 1.67, 1.40, 1.28)$. Herein, vector $\mathbf{0}$ denotes a vector of zeros; for \mathbf{a}_1 , $\mathbf{0}$ has six elements, for \mathbf{a}_2 , both $\mathbf{0}$ s have three elements each, and for \mathbf{a}_3 , $\mathbf{0}$ has six elements. Location parameters were defined as in the unidimensional model.

Population Reliability Reliability equals $\rho = (\mathbf{1}^T \Sigma_T \mathbf{1}) / (\mathbf{1}^T \Sigma_X \mathbf{1})$. Matrix Σ_T differs from matrix Σ_X by the diagonal only, which for Σ_T contains the item true-score variances. Population values of coefficients α , λ_2 , λ_4 , the *GLB*, ω_u , ω_h and ω_t (for the multidimensional model) were computed from covariance matrix Σ_X . Σ_X has diagonal item variances σ_j^2 and off-diagonal covariances σ_{jk}^2 . Following Oosterwijk et al. (2017), we compute $\sigma_j^2 = \mathcal{E}(X_j^2) - [\mathcal{E}(X_j)]^2$, and $\sigma_{jk} = \mathcal{E}(X_j X_k) - \mathcal{E}(X_j) \mathcal{E}(X_k)$. Furthermore, $\mathcal{E}(X_j) = \sum_x x P(X_j = x)$, $\mathcal{E}(X_j^2) = \sum_x x^2 P(X_j = x)$ and $\mathcal{E}(X_j X_k) = \sum_x \sum_y x y P(X_j = x, X_k = y)$, with $x = 0, \dots, 4$.

Marginal probability $P(X_j = x)$ equals

$$P(X_j = x) = \sum_{\theta_q} P(\theta_q) P(X_j = x | \theta_q), \quad (12)$$

and joint probabilities $P(X_j = x, X_k = y)$ equal

$$P(X_j = x, X_k = y) = \sum_{\theta_q} P(\theta_q) P(X_j = x | \theta_q) P(X_k = y | \theta_q). \quad (13)$$

To obtain Σ_T , we substitute the diagonal of Σ_X with the true item variances $\sigma_{T_j}^2$. We compute $\sigma_{T_j}^2$ from

$$\sigma_{T_j}^2 = \sum_{\theta_q} P(\theta_q) \left[(T_j | \theta_q) - \mathcal{E}(T_j) \right]^2, \quad (14)$$

with true scores $T_j | \theta_q = \sum_x P(X_j \geq x | \theta_q)$, and $\mathcal{E}(T_j) = \mathcal{E}(X_j)$. Probability $P(\theta_q)$ is computed as follows. First, for $Q = 1$, we compute the value of the probability density function of the standard normal distribution at each θ -value and then transform the resulting values to the zero-to-one probability scale by dividing each value by the sum of all values. Second, for $Q = 3$, the reference probability density function is the multivariate normal. Subsequently, the value of the density function is computed for each possible permutation of the three θ -vectors.

Data Generation First, for $Q = 1$, we drew N θ -values from a standard normal distribution and computed the cumulative and exact response probabilities from Eqs. (10) and (11) for each of the N θ -values. For $Q = 3$, we drew N triplets of θ -values from a multivariate normal with a specified correlation matrix ($\rho = .3$) and computed the required probabilities. Second, using the exact response probabilities,

we randomly drew N scores on sets of J Items from a multinomial distribution. We scaled these scores ordinally, consistent with a five-point Likert-scale.

Data Generation from FA-Model

We created a covariance matrix implied by a particular factor model. The procedure differed between one- and three-factor models. We used the matrix as the data-generating covariance matrix.

Unidimensional model The covariance matrix the model implies is defined as $\Sigma_U = \mathbf{c}\phi\mathbf{c}^T + \Psi$, with ϕ as the variance of factor \mathbf{g} , and Ψ as the diagonal covariance matrix of \mathbf{E} (Eq. 5; Bollen, 1989). We sampled standardized model loadings in \mathbf{c} from $U(0.3, 0.7)$. Then, the residual variances in the diagonal matrix Ψ become $1 - c^2$. We assumed factor variance is $\phi = 1$. We computed parameters for coefficients α , λ_2 , the *GLB*, and ω_u from the model-implied covariance matrix Σ_U .

Multidimensional Model The multi-factor model was a second-order model for which we assumed three group factors each explaining a unique third of the items, and one general factor explaining the group factors. Group factor loadings came from a uniform distribution $U(.4, 1)$ and loadings of group factors on the general factor came from $U(.5, 1)$. Loadings were standardized, so that squared loadings together with the residual variances added to 1. General-factor variance equaled 1. Using the Schmid-Leiman transformation (Schmid & Leiman, 1957) we transformed group and general factor loadings to loadings of a bi-factor model (Eq. 7). Residual variances were the same. The model-implied covariance matrix was $\Sigma_M = \mathbf{\Lambda}\Phi\mathbf{\Lambda}^T + \Psi$ (Bollen, 1989). Matrix $\mathbf{\Lambda}$ contains the loadings of the items on both the group factors and the general factor, Φ is the diagonal factor covariance matrix with $Q+1$ entries that equal 1, and Ψ is a diagonal matrix containing the residual variances. We computed the parameters for α , λ_2 , λ_4 , the *GLB*, ω_h , and ω_t from the model implied covariance matrix Σ_M .

Population Reliability Reliability is defined as $\rho = \sigma_T^2 / \sigma_X^2$. Assuming that squared factor loadings represent item true-score variances, population reliability equaled ω_u (see Eq. 6) for the unidimensional models and ω_t (see Eq. 9) for the multidimensional models.

Data Generation We drew random samples from a multivariate normal distribution with means of 0 and model-implied covariance matrices Σ_U and Σ_M , respectively. The resulting data were then continuous.

Factor Analysis Method We estimated all factor models by means of a confirmatory factor analysis (CFA). To obtain coefficients ω_h and ω_t , we first performed a CFA with a second-order factor model and transformed the resulting loadings into bi-factor loadings, \mathbf{c} and $\mathbf{\Lambda}$ (Eq. 7), using the Schmid-Leiman transformation.

In the simulation study, we used λ_4 as the population coefficient, and $\lambda_{4(.05)}$ as the sample estimate in the simulation runs, where $\lambda_{4(.05)}$ is the .05 quantile of a distribution of approximations to λ_4 that avoid having to consider all possible item splits, thus running into combinatorial problems (Hunt & Bentler, 2015). Coefficient $\lambda_{4(.05)}$ counteracts chance capitalization that sometimes leads to gross overestimation of population reliability.

Conditions

Numbers of items were $J = 9, 18$. Sample sizes were $N = 500, 2000$. Together with two data generation methods and two dimensionality conditions, 16 conditions resulted that were replicated 1000 times. In addition, we identified two misspecified models when estimating FA-coefficients ω_u , ω_h , and ω_t . First, we considered incorrectly assuming one factor, thus estimating coefficient ω_u , when the truth is a multi-factor model. We generated data using a second-order factor model with three group factors and a general factor, for $J = 12$ and $N = 1000$, with 1000 replications, and computed coefficient ω_u . Second, we considered incorrectly assuming multiple factors with an underlying common factor, thus computing coefficients ω_h and ω_t , when the true model is purely multi-factor and does not contain a common factor. The factor model had three orthogonal group factors. We assumed that coefficient ω_t equaled the population reliability and coefficient ω_h equaled zero as loadings on the general factor were zero. Number of items was $J = 12$, sample size was $N = 1000$, and number of replications was 1000. We estimated α , λ_2 , λ_4 , the *GLB*, ω_h , and ω_t .

Outcome Variables

We determined discrepancy, and the mean and standard error of bias. Discrepancy is the difference between parameters for reliability methods and reliability, for example, $\alpha - \rho$. Bias is the difference between the mean sample coefficient and its parameter value, for example, $\mathcal{E}(\hat{\alpha}) - \alpha$. Standard error is the standard deviation of estimates relative to the parameter, for example, $\sigma_{\hat{\alpha}} = \left(\mathcal{E}[(\hat{\alpha} - \alpha)^2] \right)^{\frac{1}{2}}$. We tested significance of the bias being different from zero.

2.2 Results

Unidimensional Models

Table 1 shows discrepancy, bias, standard error and significance results. Coefficients α , λ_2 , λ_4 , *GLB*, and ω_u showed similar results in both data generating scenarios. The discrepancy of λ_2 , the *GLB*, and ω_u was small in all unidimensional conditions.

Table 1 Discrepancy, bias, and standard error (between parentheses) of several reliability methods for unidimensional models

Coefficient	J = 9			J = 18		
	Discrepancy	N		Discrepancy	N	
		500	2000		500	2000
	IRT-data					
	$\rho = .798$			$\rho = .889$		
α	-0.48	-0.66 (0.42)	-0.32 (0.21)	-0.21	-0.34 (0.23)	-0.06 (0.12)
λ_2	-0.29	00.06 (0.41)	-0.15 (0.21)	-0.12	-0.08 (0.23)	0.05 (0.11)
λ_4	-4.46	-8.31 (0.42)*	-7.39 (0.21)*	0	-8.05 (0.22)*	-5.95 (0.12)*
<i>GLB</i>	0	28.51 (0.38)*	14.12 (0.21)*	0	-27.10 (0.19)*	13.91 (0.11)*
ω_u	-0.23	-0.21 (0.41)	-0.21 (0.21)	-0.11	-0.13 (0.23)	-0.01 (0.12)
	FA-data					
	$\rho = .748$			$\rho = .859$		
α	-4.61	-1.17 (0.53)*	-0.57 (0.27)*	-2.16	-0.46 (0.30)	-0.07 (0.15)
λ_2	-0.87	0.04 (0.52)	0.31 (0.26)*	-0.22	-0.20 (0.29)	-0.09 (0.15)
λ_4	0	-15.88 (0.53)*	-11.78 (0.28)*	0	-3.12 (0.28)*	-4.75 (0.15)*
<i>GLB</i>	0	-35.11 (0.50)*	-17.66 (0.26)*	0	33.86 (0.24)*	17.42 (0.13)*
ω_u	0	-0.47 (0.53)	-0.42 (0.26)	0	-0.13 (0.29)	0.00 (0.15)

Note. Significance is indicated with *. Table entries are transformed and rounded for better interpretation; real values are obtained by multiplying entries by 10^{-3} , e.g., the discrepancy for α ($J = 9$; IRT-data) is $-0.48 \times 0.001 = -0.00048$. Discrepancy for λ_4 was $\lambda_4 - \rho$, bias was estimated using $\lambda_{4(.05)}$

The discrepancy of coefficient λ_4 improved considerably with a larger number of items. Discrepancy was negative for all coefficients, a desirable result. Mean bias of coefficients α , λ_2 , and ω_u was relatively small. Table 1 shows that the discrepancy of the *GLB* is almost equal to 0, but its bias is largely positive, a finding consistent with results reported by Oosterwijk et al. (2017). Estimate $\lambda_{4(.05)}$ underestimated population value, λ_4 . Increase in sample size resulted in better performance for all coefficients. Except for the *GLB*, an increase in the number of items led to smaller bias. Except for λ_4 and the *GLB*, the coefficients' performance was satisfactory across all unidimensional conditions.

Multidimensional Models

Table 2 shows that all coefficients had smaller discrepancy and bias as samples grew larger and, except for the *GLB*, results improved as the number of items grew. Discrepancy was highly similar for both data generation procedures. As expected, discrepancy of lower bounds α and λ_2 was much larger for the multidimensional data than for the unidimensional data. Coefficient λ_4 showed an unexpectedly large discrepancy with the multidimensional IRT-data and considerable negative bias throughout all multidimensional conditions. The *GLB* had very small discrepancy but an expectedly large bias. As expected, an increase in the number of items

Table 2 Discrepancy, bias, and standard error (between parentheses) of several reliability methods for multidimensional models

Coefficient	J = 9			J = 18		
	Discrepancy	N		Discrepancy	N	
		500	2000		500	2000
	IRT-data					
	$\rho = .738$			$\rho = .853$		
α	-82.13	-0.38 (0.73)	-0.18 (0.36)	-44.61	-0.77 (0.39)*	-0.39 (0.20)
λ_2	-67.82	-1.41 (0.65)*	-0.24 (0.32)	-36.14	-0.23 (0.35)	-0.12 (0.18)
λ_4	-31.74	-16.32 (0.59)*	-14.94 (0.31)*	-0.04	-13.72 (0.31)*	-8.65 (0.16)*
<i>GLB</i>	-0.07	26.15 (0.52)*	-12.08 (0.28)*	0	32.34 (0.25)*	16.37 (0.15)*
ω_h	-325.91	7.01 (1.39)*	1.82 (0.69)*	-376.23	3.41 (1.23)*	-0.18 (0.64)
ω_t	-0.10	0.70 (0.55)	0.05 (0.27)	-0.04	-0.37 (0.29)	-0.24 (0.15)
	FA-data					
	$\rho = .872$			$\rho = .917$		
α	-59.19	-1.46 (0.40)*	-0.59 (0.21)*	-29.84	-0.36 (0.23)	-0.20 (0.12)
λ_2	-49.93	-0.67 (0.36)	-0.37 (0.19)*	-24.41	-0.10 (0.22)	-0.06 (0.11)
λ_4	-13.18	-30.18 (0.34)*	-26.8 (0.17)*	-0.28	-11.68 (0.17)*	-11.99 (0.08)*
<i>GLB</i>	0	11.74 (0.26)*	-5.65 (0.14)*	0	-18.14 (0.15)*	9.17 (0.08)*
ω_h	-217.61	-0.98 (0.83)	-0.87 (0.42)*	-212.22	-0.42 (0.69)	-0.20 (0.35)
ω_t	0	-0.55 (0.27)*	-0.24 (0.14)	0	-0.09 (0.17)	-0.04 (0.08)

Note. Significance is indicated with *. Table entries are transformed and rounded for better interpretation; real values are obtained by multiplying entries by 10^{-3} , e.g., the discrepancy for α ($J = 9$; IRT-data) is $-82.1 \times 0.001 = -0.0821$. Discrepancy for λ_4 was $\lambda_4 - \rho$, bias was estimated using $\lambda_{4(.05)}$

produced a larger bias for the *GLB*, because capitalization on chance increases with the number of items.

Results for coefficient ω_h were not consistent with results for the other estimators. The population value of the coefficient ω_h was much lower than the population reliability. This was expected, given that ω_h indicates how well a common attribute is represented irrespective of the real factor structure. The difference between coefficients ω_h and ω_t indicates the presence of multidimensionality. Coefficient ω_h performed well with the FA-data, but with the IRT-data bias was positive, meaning it overestimated the population ω_h .

Coefficient ω_t performed well across all multidimensional conditions. It had negligible discrepancy (by definition, zero with the FA-data) and small mean bias across all conditions.

Misspecified Models

In the first case, the misspecification occurred by estimating the wrong coefficient, ω_u , which is suited for one-factor data expect data were in fact multi-factorial. The population value of coefficient ω_u was far from the population reliability (Table 3).

Table 3 Discrepancy, bias, and standard error (between parentheses) of several reliability methods for misspecified models

Coefficient	Discrepancy	Bias
	Case (1), $\rho = .899$	
ω_u	-42.18	-0.80 (0.22)*
	Case (2), $\rho = .807$	
α	-155.43	-1.23 (0.52)*
λ_2	-102.09	-0.15 (0.37)
λ_4	-2.95	-25.57 (0.32)*
<i>GLB</i>	0	19.88 (0.27)*
ω_h	-805.02	155.67 (3.76)*
ω_t	0	0.13 (0.27)

Note. Case (1): Multi-factor population model and data; computations assumed unidimensionality. Case (2): Multi-factor population model with group factor but no common factor, and data; computations assumed a common factor. Significance is indicated with a *. Table entries were transformed and rounded for better interpretation; real values are obtained by multiplying entries by 10^{-3} , e.g., discrepancy for ω_u is $-42.18 \times 0.001 = -0.04218$. $J = 12$ items and $N = 1000$

Subsequently, the estimates for ω_u were far off. In the second case, multi-factor data with a common factor was incorrectly assumed when the model generating the data contained only group factors but no common factor. The discrepancy of coefficients α and λ_2 was quite large, mirroring the multidimensionality of the data (Table 3). Coefficient λ_4 and the *GLB* had small discrepancy. The discrepancy of coefficient ω_h was huge, meaning the coefficient properly identified the absence of a common attribute. Because data were noisy, ω_h had considerable bias. Coefficient ω_t showed small bias. Its discrepancy was 0 since we used a factor model to generate the data. Arguably, a cautious researcher should always check model fit before estimating reliability coefficients that assume a certain structure of the data.

3 Discussion

The population values of the reliability methods were all fairly close to the population reliability with unidimensional data, which changed when data were multidimensional. Most coefficients had small bias in almost all conditions, except for the positive bias of the *GLB* and the negative bias of λ_4 . All coefficients did well with unidimensional data. For multidimensional data, coefficients α and λ_2 on average underestimated reliability, while other methods were closer to reliability. For high reliability, coefficients α and λ_2 had high values, albeit somewhat smaller than the true reliability. The question is whether in this situation one should rather estimate reliability for each dimension separately. Among other things, this depends

on the practical use of the test when it is sensible to distinguish different attributes or different aspects of the same attribute. In general, FA-coefficients performed very well. Coefficient λ_4 and the corrected estimate $\lambda_{4(.05)}$ were not satisfactory, because the discrepancy between λ_4 and population reliability was often larger than expected and the bias of $\lambda_{4(.05)}$ was too large to distinguish the coefficient from other lower bounds such as α .

The goal of the simulation study was investigating whether the reliability methods performed differently with discrete data generated by IRT and continuous data generated by FA. Does this difference in data types cause problems when comparing reliability methods results? We argue it does not, because different data types only present another hurdle the coefficients have to take rendering their performance evaluation more interesting. Regarding our simulation outcomes (discrepancy and bias), we found that the methods performed equally well with ordinal IRT data as with continuous FA data.

A limitation to our study was that we considered point estimation, but interval estimation, which is not common practice yet, may be more informative. Recent studies have shown that with unidimensional and multidimensional data, Bayesian credible intervals for coefficients α , λ_2 , the GLB, ω_u , ω_h , and ω_t perform well (Pfadt, van den Bergh, & Moshagen, 2021a, b). We assume that the credible intervals of the reliability coefficients relate to population reliability in the same way as the population values of the coefficients do (as denoted by the discrepancy values we found).

In addition to the dominant CTT and FA reliability methods, less well-known methods based on IRT (Holland & Hoskens, 2003; Kim, 2012) and generalizability theory (GT; e.g., Brennan, 2001) exist. In IRT, use of typical IRT reliability methods is rare given the focus on the scale-dependent information function, which proves to be a powerful tool in IRT applications, such as adaptive testing and equating. GT provides an attempt to incorporate the influence of different facets of the test design and environment in the estimation of reliability. Suppose one studies the effect of test version and score rater on test performance. This requires a design with factors persons (i), test versions (t), and raters (r). Item scores are decomposed into person effect (v_i), test effect (v_t), and rater effect (v_r), interaction effects, and a residual effect (Δ_{itr}), comparable with an ANOVA layout, so that $X_{itr} = \mu + v_i + v_t + v_r + v_{it} + v_{ir} + v_{tr} + \Delta_{itr}$. Reliability methods, called generalizability and dependability methods identify variance sources that affect relative and absolute person ordering, respectively, and correct for other, irrelevant sources. The GT approach provides a different perspective relevant to some research contexts where richer data are available and is worth pursuing in future research.

To conclude, when researchers have unidimensional data, the choice of a reliability coefficient is mostly arbitrary (if λ_4 and the *GLB* are discarded). With multidimensional data, the use of a factor model coefficient is encouraged, but lower bounds such as α prevent researchers from being too optimistic about reliability.

References

- Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons, Inc.. <https://doi.org/10.1002/9781118619179>
- Brennan, R. L. (2001). *Generalizability theory*. Springer. <https://doi.org/10.1007/978-1-4757-3456-0>
- Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, 55(4), 577–601. <https://doi.org/10.1007/BF02294609>
- Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika*, 68(1), 123–149. <https://doi.org/10.1007/BF02296657>
- Hunt, T. D., & Bentler, P. M. (2015). Quantile lower bounds to reliability based on locally optimal splits. *Psychometrika*, 80(1), 182–195. <https://doi.org/10.1007/s11336-013-9393-6>
- Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36(2), 109–133. <https://doi.org/10.1007/BF02291393>
- Kim, S. (2012). A note on the reliability coefficients for item response model-based ability estimates. *Psychometrika*, 77(1), 153–162. <https://doi.org/10.1007/s11336-011-9238-0>
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- McDonald, R. P. (1999). *Test theory: A unified treatment* (1st ed.). Psychology Press. <https://doi.org/10.4324/9781410601087>
- Oosterwijk, P. R., Van der Ark, L. A., & Sijtsma, K. (2017). Overestimation of reliability by Guttman's λ_4 , λ_5 , and λ_6 and the greatest lower bound. In L. A. van der Ark, S. Culppepper, J. A. Douglas, W.-C. Wang, & M. Wiberg (Eds.), *Quantitative psychology research: The 81th annual meeting of the psychometric society 2016* (pp. 159–172). Springer. https://doi.org/10.1007/978-3-319-56294-0_15
- Pfadt, J. M., van den Bergh, D., Moshagen, M. (2021a). *The reliability of multidimensional scales: A comparison of confidence intervals and a Bayesian alternative* [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/d3gfs>
- Pfadt, J. M., van den Bergh, D., Sijtsma, K., Moshagen, M., & Wagenmakers, E. -J. (2021b). Bayesian estimation of single-test reliability coefficients. *Multivariate Behavioral Research*. <https://doi.org/10.1080/00273171.2021.1891855>
- Samejima, F. (1968). *Estimation of latent ability using a response pattern of graded scores*. Educational Testing Service.
- Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, 22(1), 53–61. <https://doi.org/10.1007/BF02289209>
- Sijtsma, K., & Pfadt, J. M. (2021). Part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrika*, 86(4), 843–860. <https://doi.org/10.1007/s11336-021-09789-8>
- Sijtsma, K., & Van der Ark, L. A. (2021). *Measurement models for psychological attributes* (1st ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9780429112447>
- Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1), 72–101. <https://doi.org/10.2307/1412159>
- Woodhouse, B., & Jackson, P. H. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: II: A search procedure to locate the greatest lower bound. *Psychometrika*, 42(4), 579–591. <https://doi.org/10.1007/bf02295980>
- Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's alpha, Revelle's beta, and McDonald's omega h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70(1), 123–133. <https://doi.org/10.1007/s11336-003-0974-7>
- Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for omega h. *Applied Psychological Measurement*, 30(2), 121–144. <https://doi.org/10.1177/0146621605278814>

Modeling Covarying Responses in Complex Tasks



Amanda Luby and Riley E. Thompson

Abstract In testing situations, participants are often asked for supplementary responses in addition to the primary response of interest, which may include quantities like confidence or reported difficulty. These additional responses can be incorporated into a psychometric model either as a predictor of the main response or as a secondary response. In this paper we explore both of these approaches for incorporating participant's reported difficulty into a psychometric model using an error rate study of fingerprint examiners. Participants were asked to analyze print pairs and make determinations about the source, which can be scored as correct or incorrect decisions. Additionally, participants were asked to report the difficulty of the print pair on a five point scale. In this paper, we model (a) the responses of individual examiners without incorporating reported difficulty using a Rasch model, (b) the responses using their reported difficulty as a predictor, and (c) the responses and their reported difficulty as a multivariate response variable. We find that approach (c) results in more balanced classification errors, but incorporating reported difficulty using either approach does not lead to substantive changes in proficiency or difficulty estimates. These results suggest that, while there are individual differences in reported difficulty, these differences appear to be unrelated to examiners' proficiency in correctly distinguishing matched from non-matched fingerprints.

Keywords Item response theory · Forensic science · Bayesian statistics

This work was partially funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreements 70NANB15H176 and 70NANB20H019 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, Duke University, University of California Irvine, University of Virginia, West Virginia University, University of Pennsylvania, Swarthmore College and University of Nebraska, Lincoln.

A. Luby (✉) · R. E. Thompson
Department of Mathematics and Statistics, Swarthmore College, Swarthmore, PA, USA
e-mail: aluby1@swarthmore.edu

1 Introduction

It is increasingly common to collect collateral responses alongside the main responses of interest in testing or survey situations. For example, we may record a confidence level for a multiple-choice question, response time for a Likert-scale survey question, or ask participants to report how difficult they found an item. In this paper, we outline two general approaches for incorporating this type of observed data into a psychometric model:

1. as a covariate for predicting responses
2. as a second response using additional latent variables.

In the first setting, the covarying responses are treated as *predictors* for our outcome variable of interest, so we would use confidence or response time to explain our primary response variable (usually correctness). In the second setting the covarying responses are treated as a second *outcome*, so we would model confidence or response time alongside correctness as a multivariate response variable.

Throughout this paper, we will use Y_{ij} to denote our main outcome of interest, where i indexes participants and j indexes items. X_{ij} represents the covarying response. In approach (1), we model $P(Y_{ij}|\theta_i, b_j, X_{ij})$, while in the second approach we model, $P(Y_{ij}, X_{ij}|\theta_i, b_j, \psi)$, where θ_i represents participant i 's proficiency, b_j represents item j 's difficulty, and ψ represents possible additional latent variables. These two approaches and their uses in the literature are discussed in further detail in Sect. 2.

We demonstrate the use of these two approaches and compare their performance using data collected in the FBI ‘‘Black Box’’ study (Ulery et al., 2011). The purpose of this study was to evaluate the accuracy and reliability of decisions made by fingerprint examiners in the U.S. After fingerprint examiners completed each question, they were also asked to report the difficulty of the comparison on a 5-point scale. Reported difficulty is likely to vary depending on both the examiner and the item, and may provide further information about θ , b , or allow for the estimation of further latent variables. Luby (2019) demonstrated the use of item response models such as the Rasch (1960) and partial credit (Masters, 1982) models on this data, Luby et al. (2020) applied more advanced process models and used the joint modeling approach outlined above to account for the reported difficulty, and Luby et al. (2021) expanded on the use of tree-based modeling approaches for this data (De Boeck & Partchev, 2012; Jeon et al., 2017). In the current paper, we evaluate the use of the joint modeling approach compared to a covariate approach.

The remainder of the paper is structured as: Sect. 2 introduces the modeling framework we use throughout, Sect. 3 discusses the fingerprint comparison application in further detail, and in Sect. 4 the results are presented. We discuss limitations and future work in Sect. 5.

2 Methods

2.1 Rasch Model

The Rasch model (Rasch, 1960; Fischer & Molenaar, 2012), a simple yet powerful IRT model, uses a mixed effect logistic regression approach with random effects for both participant proficiency, θ_i , and item difficulty, b_j . Responses can be represented in an $I \times J$ matrix. The information stored within the matrix is a binary response variable, where a 1 corresponds to a correct response and a 0 corresponds to an incorrect response. Below is an example of such a matrix:

$$Y = \begin{bmatrix} 1 & 0 & - & \dots & 1 \\ 0 & - & 1 & \dots & 0 \\ 1 & 1 & - & \dots & - \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & - & \dots & 1 \end{bmatrix}.$$

The probability of a correct response is formulated as in Eq. 1:

$$P(Y_{ij} = 1) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)}. \quad (1)$$

Note that additional constraints are needed in order to identify the model. Using a Bayesian modeling approach, we use the convention of setting the mean of the proficiency parameters to be zero. We also use the recommended priors for efficiency in estimation in the `brms` R package (Bürkner, 2017). That is,

$$\left. \begin{aligned} Y_{ij} &\sim \text{Bernoulli}(p_{ij}) \\ p_{ij} &= \text{logit}^{-1}(\theta_i - b_j) \\ \theta_i &\sim N(0, \sigma_\theta^2) \\ b_j &\sim N(0, \sigma_b^2) \\ \sigma_\theta, \sigma_b &\sim \text{Half-T}_3(0, 2.5^2), \end{aligned} \right\} \quad (2)$$

where $\text{Half-T}_3(0, 2.5^2)$ refers to a folded T-distribution with 3 degrees of freedom.

Using `brms`, it is possible to fit Bayesian multilevel models using identical syntax to the `lme4` package (Bates et al., 2015), which fits multilevel models using maximum likelihood techniques. `brms` uses the Stan modeling language (Stan Development Team, 2018b) for estimation and interfaces with R (R Core Team, 2013; Stan Development Team, 2018a). Stan is a probabilistic programming language for statistical inference that fits Bayesian models using Hamiltonian Monte Carlo and No-U-Turn sampling. For further information on fitting IRT models with `brms`, see Bürkner (2019) for an excellent overview.

2.2 Rasch Model with Response-Level Covariates

There is a rich literature on a number of different approaches for incorporating variables that are not the response variable into an IRT analysis. When the additional variables are covariates describing either the participants or the items, they can be used as predictors for proficiency or difficulty in the IRT model (De Boeck & Wilson, 2004). A further approach is Differential Item Functioning (DIF), which represents the additional covariate as an interaction between an item indicator and a person predictor representing group membership (Holland & Wainer, 2012).

The first method we use to incorporate a covarying response is to treat it as a covariate at the response level. That is, we do not consider it as a covariate specific to participants or items, but one that varies with both. Similar to the response matrix above, we can represent the covarying responses as an $i \times j$ matrix, where each row corresponds to a participant's responses and each column corresponds to an item. In the case where the covarying response is an ordered categorical variable, such as confidence or reported difficulty, with ordered categories $A < B < C < D < E$, this matrix will be something like the following:

$$X = \begin{bmatrix} A & C & - & \dots & B \\ C & - & C & \dots & C \\ D & A & - & \dots & - \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ E & B & - & \dots & A \end{bmatrix} \tag{3}$$

Let the covarying response matrix X be as above, and let \mathbf{X}_{ij} be a vector of indicator variables for each of the possible categories A-E. We can then represent the Rasch model with Response-Level covariates as in Eq. 4:

$$\left. \begin{aligned} Y_{ij} &\sim \text{Bernoulli}(p_{ij}) \\ p_{ij} &= \text{logit}^{-1}(\theta_i - b_j + \boldsymbol{\beta}\mathbf{X}_{ij}) \\ \theta_i &\sim N(0, \sigma_\theta) \\ b_j &\sim N(0, \sigma_b) \\ \beta_k &\sim N(0, 5) \\ \sigma_\theta, \sigma_b &\sim \text{Half-T}_3(0, 2.5^2). \end{aligned} \right\} \tag{4}$$

However, this exact model formulation may not be appropriate for every type of covarying response. In the case of response time, for example, the covarying response matrix could instead contain real-valued responses, in which case the indicator variables defined above would not be needed. Care should be taken when choosing prior distributions to ensure that the type of covarying response is accounted for.

2.3 Joint Rasch Model

Variables such as confidence or response time, which are not observed prior to observing the responses, could not be used to predict responses in a realistic setting. Using covarying responses as a predictor in the model for responses may improve predictive performance of a model, but are not truly a ‘predictor’. While this is not necessarily an issue for inferential modeling, it does suggest an alternative framework may be more appropriate.

As an alternative, we can treat the covarying response as a secondary outcome using a multivariate response. Assuming the same X and Y response matrices defined above, the joint responses model we use is given in Eqs. 5 and 6.

$$\left. \begin{aligned} Y_{ij} &\sim \text{Bernoulli}(p_{ij}) \\ p_{ij} &= \text{logit}^{-1}(\theta_i - b_j) \\ X_{ij}^* &= \text{logit}^{-1}(\beta_0 + h_i + f_j) \\ (\theta_i, h_i) &\sim \text{MVN}(0, \sigma_\theta L_\theta L_\theta' \sigma_\theta) \\ (b_j, f_j) &\sim \text{MVN}(0, \sigma_b L_b L_b' \sigma_b) \\ L_\theta, L_b &\sim \text{LKJ}(1) \\ \beta_0, \gamma_i &\sim N(0, 5) \\ \sigma_\theta, \sigma_b &\sim \text{Half-T}_3(0, 2.5^2) \end{aligned} \right\} \quad (5)$$

$$X_{ij} = \begin{cases} \text{A} & X_{ij}^* \leq \gamma_1 \\ \text{B} & \gamma_1 < X_{ij}^* \leq \gamma_2 \\ \text{C} & \gamma_2 < X_{ij}^* \leq \gamma_3 \\ \text{D} & \gamma_3 < X_{ij}^* \leq \gamma_4 \\ \text{E} & X_{ij}^* > \gamma_4. \end{cases} \quad (6)$$

To model the ordered categorical variable X_{ij} , a continuous latent variable for each (i, j) pair, X_{ij}^* , is estimated and then binned into the categorical variable X_{ij} according to category cutoffs γ_i . We additionally constrain $\gamma_1 < \gamma_2 < \gamma_3 < \gamma_4$ to account for the ordered categories. The additional variables h_i and f_j allow for the possibilities that participants over- or under-estimate the covarying response relative to the reporting tendencies of other participants. The f_j term tells us whether a similar difference is present for the items. To estimate θ, b, h and f , we assume:

1. $Y_i \perp Y_{i'}$: participant responses are independent of other participants’ responses
2. $Y_{ij} \perp Y_{ij'} | \theta_i$ and $X_{ij} \perp X_{ij'} | \theta_i, h_i$: participant responses (and covarying responses) are conditionally independent given the latent variables θ, h for any given participant
3. $X_{ij} \perp Y_{ij} | \theta_i, b_j, g, h_i, f_j$: covarying responses are conditionally independent of primary responses given all latent variables for each i, j pair.

Assumptions (1) and (2) correspond to the local independence assumption and are common in the IRT literature. Assumption (3) is an extension of (2) to the covarying response variable.

Thissen (1983) provides an early example of this type of modeling, where the logarithm of response time is modeled as a linear function of the standard IRT quantity $(\theta_i - b_j)$ and additional latent variables for both items and participants. Ferrando and Lorenzo-Seva (2007), van der Linden (2006) each propose various models for modeling response time jointly with the traditional correct/incorrect IRT response. Modeling collateral information alongside responses in such a way has been shown to improve estimates of IRT parameters through the sharing of information (van der Linden et al., 2010).

2.4 Model Evaluation

We use three approaches to evaluate model fit. First, a posterior predictive check provides a high-level measure of whether the predicted responses follow a similar distribution as the observed responses. We then assess in-sample predictive performance for both correct and incorrect responses. Finally, we estimate the out-of-sample predictive performance using the widely-applicable information criterion (WAIC, Watanabe, 2010). Each of these evaluation metrics is discussed in more detail below.

Posterior Predictive Check For a preliminary assessment of model validity, we perform a posterior predictive check (Sinharay et al., 2006). In such a check, parameter values from each posterior draw are used to generate a simulated dataset, resulting in a large number of replicate datasets. We then compute the predicted responses (Y_{ij}) for each $i \times j$ pair in the original dataset and display the overall distribution aggregated over all responses. If the model is performing adequately, the distribution of the simulated responses will be similar to the observed responses.

In-sample Predictive Performance We also assess model fit using in-sample predictive performance. While the posterior predictive check above investigates the overall distribution of predicted responses, it does not measure whether the model is correctly predicting individual observations. The in-sample predictive performance investigates how well the model is predicting responses at the observation level.

In this check, we use the posterior means to compute $P(Y_{ij} = 1)$ under the model for each $i \times j$ pair. If $P(Y_{ij} = 1) > 0.5$, we predict a correct response, otherwise we predict an incorrect response. We then compare each predicted response to the original response, and summarize model performance through the number of false positive errors (i.e. the model predicts correct when the observed response was incorrect) and false negative errors (i.e. the model predicts incorrect when the observed response was correct). If the number of false positive and false negative errors is low, the model is performing well in terms of predicted performance.

Out-of-sample Predictive Performance To complement the in-sample predictive performance described above, we also estimate the out-of-sample predictive performance. Out-of-sample predictive performance describes how well the model performs at prediction, while correcting for the bias in using the same set of data to evaluate the model that was used to estimate the parameters. We use the widely-applicable information criterion (WAIC, Watanabe, 2010) as defined in Vehtari et al. (2017) to compare the model fit. This quantity provides an estimate of pointwise out-of-sample predictive performance using posterior draws:

$$\text{WAIC} = -2 \times (\widehat{\text{lpd}} - \sum_{n=1}^N V_{s=1}^S \log(p(y_n, x_n | \theta^s, b^s, f^s, h^s))) \quad (7)$$

where $V_{s=1}^S$ represents the sample variance, θ^s, b^s, f^s, h^s are the parameter estimates from the s th posterior draw, and

$$\widehat{\text{lpd}} = \sum_{n=1}^N \log \frac{1}{S} \sum_{s=1}^S p(y_n, x_n | \theta^s, b^s, f^s, h^s)$$

is the log pointwise predictive density. Estimates of the standard error of the WAIC ($\widehat{SE}(\text{WAIC})$) can also be computed. If the WAIC of one model is lower (better) than a second model, but the two WAIC values are within two standard errors from one another, the difference may be due to sampling error and not to a true difference in model performance (Vehtari et al., 2017).

Parameter Estimates In addition to model fit assessments, we also use the parameter estimates to gauge the performance of each model. We look for reasonable population-level covariates and informative latent variables describing the participants. By ‘population-level’ covariates, we mean that these covariates are not specific to participants or to items but instead describe the responses at the $i \times j$ level (e.g. the β estimates from Eq. 8 and 9). We also investigate the correlation between the latent variables across items and participants. If the correlations are strong, the reported difficulty responses likely measure similar trends as the comparison responses. Weak correlations, on the other hand, suggest that the two responses are measuring different latent tendencies.

3 Application: Forensic Fingerprint Comparisons

For the remainder of the paper, we will demonstrate each of the modeling frameworks introduced above applied to forensic fingerprint comparisons. The general task involves comparing a crime-scene print (called a *latent* print) to a known-source *reference* print to determine whether the two prints came from the same source.

ACE-V Procedure The process in which examiners compare prints is called ACE-V (Analysis, Comparison, Evaluation, Verification). In the analysis stage, examiners are shown a latent print and are tasked with determining if the print has value for individualization (VID), value for exclusion only (VEO) or no value (NV).¹ If a print has value, either VEO or VID, the examiner is then presented with a reference print in the comparison stage. Prints with no value are no longer examined and do not proceed into the comparison stage.

In the comparison stage, an examiner compares the latent and reference print side by side. This usually involves comparing the ‘minutiae’, or small details in the fingerprint ridge pattern, between the two prints. The evaluation stage occurs when the examiner determines whether or not the reference and latent print are from the same source. In the evaluation stage, an examiner can make one of three determinations: individualization, exclusion, or inconclusive.² ‘Individualization’ means the examiner concluded that the two prints are from the same source, exclusion means they concluded that the prints are from two difference sources, and inconclusive comparisons mean that the examiner could not determine whether or not the two prints are from the same source. Verification requires independent confirmation of the determination from another fingerprint examiner. For more details on fingerprint comparison and the ACE-V process, see AAAS (2017).

FBI Black Box Study The FBI, with NOBLIS Inc., conducted the Black Box study to simulate realistic casework for finger print examiners (Ulery et al., 2011). The purpose of the Black Box Study was to analyze the accuracy and reliability of fingerprint examiners. Specifically, the goal of the study was to calculate the frequency of false positive and false negative responses. A false positive occurs when an examiner individualizes non-mated pairs and a false negative occurs when an examiner excludes mated pairs.

A total of 169 latent print examiners participated in the study. Each examiner was given a randomized examiner ID to keep their responses anonymous. Examiners were assigned a random set of roughly 100 latent prints to analyze using the ACE-V framework. There were 365 latent prints and 484 reference prints from 21 people, which were combined to form 744 latent-reference pairs. Each latent-reference pair was assigned a pair ID.

A mated pair describes two prints in which the prints were obtained from the same source. Therefore, for any mated pair both the latent and exemplar print are from the same finger on the same person. Non-mated pairs, however, are two prints

¹ These latent evaluation categories may vary depending on different laboratory practices. We use the categories that were recorded in the Black Box study (Ulery et al., 2011).

² Individualizations are no longer recommended in practice, in favor of ‘identification’ or ‘same source’ conclusions. Since the data used in this paper was collected in 2011 and used the ‘Individualization’ terminology, this is what we use throughout. See Friction Ridge Subcommittee of the Organization of Scientific Area Committees for Forensic Science (2017, 2019) for further discussion and current recommendations.

that are not from the same source. This means that the latent and exemplar print are from different people or different fingers on the same person.

For the purposes of this paper, correct responses were defined to be individualizing mated pairs and excluding non-mated pairs. All inconclusive responses and well as erroneous responses, such as individualizing non-mated pairs and excluding mated pairs, were scored as incorrect.

Covarying Response: Reported Difficulty Reported difficulty was recorded for each complete response in the Black Box study. A ‘complete’ response refers to the examiner making it to the comparison/evaluation stage and making a source decision (individualization vs. exclusion vs. inconclusive). After examiners came to a decision, they were asked to report the difficulty of the comparison task. The difficulty scale examiners were asked to report on was: A-Obvious, B-Easy, C-Medium, D-Difficult, or E-Very Difficult. See Table 1 for descriptions of the categories that were given to participants while completing the Black Box study, as well as the proportions of each response.

We emphasize two reasons that interpreting reported difficulty is a complex task. First, individuals may naturally have subjective thresholds on what constitutes a ‘difficult’ as compared to a ‘very difficult’ item, for example. These subjective thresholds may also be influenced by the type of training or typical casework involved in that particular examiner’s career. Second, reported difficulty itself is ambiguous. For example, in the Analysis stage an examiner can determine that a latent print has value (VID or VEO), but make a determination of Inconclusive. An examiner could have quickly and “easily” came to the inconclusive determination and rated the difficulty either A-Obvious or B-Easy. This scenario differs from one in which an examiner was able to individualize mates or exclude non-mates with ease.

Table 1 shows the infrequency in which examiners select difficulty ratings ‘Obvious’ and ‘Very Difficult’. A simple solution is to regroup the original difficulty scale to have a more even distribution of responses per difficulty. The original difficulty scale was rescaled such that ‘Obvious’ and ‘Easy’ were combined to the category entitled ‘Easy’. Additionally, ‘Difficult’ and ‘Very Difficult’ were grouped

Table 1 Proportion of responses for reported difficulty category, along with the description provided to participants while completing the Black Box study

Difficulty rating	Proportion	Description
Obvious	8.14%	The comparison determination was obvious.
Easy	26.45%	... easier than most latent comparisons.
Moderate	44.31%	... a typical latent comparison.
Difficult	17.87%	... more difficult than most latent comparisons
Very difficult	3.23%	... unusually difficult, involving high distortion and/or other red flags

Table 2 Proportion of responses for each transformed category, including corresponding original difficulty categories

Transformed difficulty	Original difficulty	Proportion
Easy	Obvious/easy	34.59%
Medium	Moderate	44.31%
Hard	Difficult/very difficult	21.10%

into the category ‘Hard’. The 3-point scale created more balance amongst the new three categories (see Table 2).

A second reason we might prefer the transformed difficulty scale is that responses appear to be more stable across time. In a follow up study in which the set of examiners re-analyzed 25 pairs that they had seen in the original Black Box study (Ulery et al., 2012), ‘Medium’ was the only category in which a majority of responses did not change: 57.35% of comparisons that were reported to be ‘A-Obvious’ on the original Black Box study changed in the retest study, 51.63% of original ‘B-Easy’ responses changed, 45.27% of original ‘C-Medium’ responses changed, 58.26% of ‘D-Difficult’ responses changed, and 100% of original ‘E-Very Difficult’ responses changed in the retest study. By combining ‘A-Obvious’ with ‘B-Easy’, and ‘E-Very Difficult’ with ‘D-Difficult’, responses become more stable. On the transformed scale, only 37.20% of responses that were ‘Easy’ (‘A-Obvious’ and ‘B-Easy’ combined) on the initial study changed to a different difficulty on the retest, 45.27% of ‘Medium’ responses changed, and 51.31% of ‘Hard’ (‘D-Difficult’ and ‘E-Very Difficult’ combined) changed on the retest study. While we still observe a substantial proportion of responses changing on the retest, the transformed 3-category difficulty scale appears to be more stable across time than the original 5-category difficulty scale.

3.1 Model Formulation

Rasch Model The first model that we fit to the Black Box data was the standard Rasch model as outlined in Sect. 2, Eq. 1, which estimates a proficiency for each participant (θ_i) and a difficulty for each item (b_j).

Covariate (5-Point Scale) The reported difficulty scale can be added to the Rasch model as a population-level covariate. This formulation results in coefficients that are constant across all responses. In other words, the difficulty coefficient does not vary depending on the item or the examiner. The probability of a correct response using the covariate approach can be modeled as:

$$P(Y_{ij} = 1) = \text{logit}^{-1}(\theta_i - b_j + \beta \mathbf{D}_5) \quad (8)$$

where \mathbf{D}_5 consists of indicator variables for each of the five possible reported difficulties representing the examiner’s difficulty response on the pair.

Covariate (3-Point Scale) The transformed difficulty scale can also be added to the Rasch model as a population-level covariate. The probability of a correct response using the covariate approach can be modeled as:

$$P(Y_{ij} = 1) = \text{logit}^{-1}(\theta_i - b_j + \beta \mathbf{D}_3) \quad (9)$$

where \mathbf{D}_3 consists of indicator variables for each of the three possible rescaled difficulties for each examiner's response on the item.

Joint Response (5-Point Scale) In the joint response approach (outlined in more detail in Sect. 2), we treat both Y_{ij} (correctness) and X_{ij} (reported difficulty) as outcome variables:

$$P(Y_{ij} = 1) = \text{logit}^{-1}(\theta_i - b_j) \quad (10)$$

$$X_{ij}^* = \text{logit}^{-1}(\beta_0 + h_i + f_j) \quad (11)$$

and X_{ij} consists of the reported difficulty responses on the original five-point scale and is modeled with the piecewise cutoff function in Eq. 6.

Joint Response (3-Point Scale) The joint response model for the 3-point scale is formulated the same as the 5-point scale (See Eqs. 9 and 10), but X_{ij} consists of the rescaled difficulty responses on the three-point scale and is modeled with a modified piecewise cutoff function as in Eq. 6, with three possible categories.

4 Results

In this section, we focus on two aspects of the results of the model. First, we present a comparison of the model fits using the methods outlined in Sect. 2. Second, we investigate the latent parameters for participants (θ and h) across each of the five models, their correlation, and the estimated impact of reported difficulty.

4.1 Model Comparison

Posterior Predictive Check We begin by performing a simple posterior predictive check on all five models. In Fig. 1, we compare the overall number of predicted incorrect ($y_{rep} = 0$) and correct ($y_{rep} = 1$) among each of the posterior draws to the actual number of correct and incorrect responses observed in the Black Box study. We see that all models do quite well at matching the overall distribution of responses. The same type of posterior predictive check was also performed for the covarying responses in the Joint Responses models, and the models again performed quite well at capturing the overall number observed in each category (see Fig. 2).

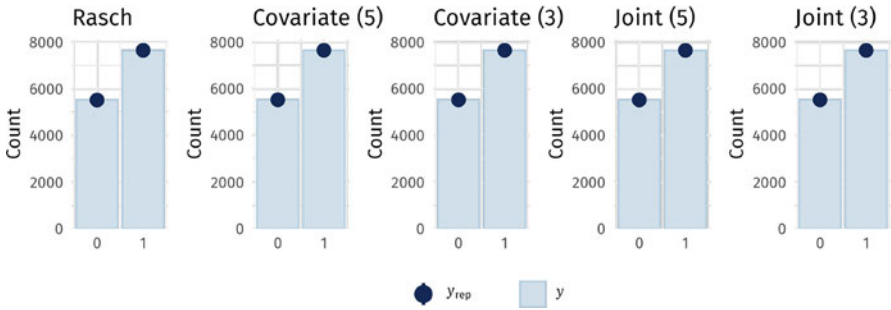


Fig. 1 Results from a posterior predictive check for the Rasch model, Covariate models (both 5-category and transformed 3-category), and joint models (both 5-category and transformed 3-category). Simulated responses based on posterior draws are denoted y_{rep} , while observed responses are denoted by y . All five models perform quite well according to the posterior predictive check

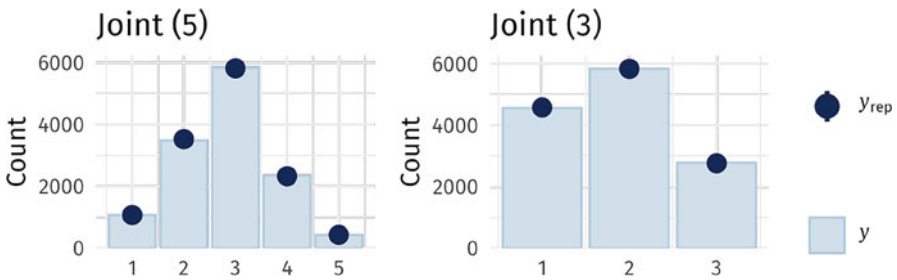


Fig. 2 Results from a posterior predictive check with reported difficulty as the response for the two joint models. Simulated responses based on posterior draws are denoted y_{rep} , while observed responses are denoted by y . Both joint response models perform quite well according to the posterior predictive check

In-sample Predictive Performance In-sample predictions for each Y_{ij} observation, for each of the five models, is displayed in Fig. 3. In general, each model tends to make more “false positive” predictions (observed incorrect but $P(\text{Correct}) > 0.5$) than “false negative” predictions (observed correct but $P(\text{Correct}) < 0.5$). While different thresholds for $P(\text{Correct})$ could also be used to optimize model performance, our primary interest is comparing models to one another and so we use 0.5 for convenience and interpretability.

We also observe that the joint models (“Joint (5)” and “Joint (3)”) are more balanced in the error rates compared to models that include reported difficulty as a covariate (“Covariate (3)” and “Covariate (5)”): they tend to make fewer false positive errors and more false negative errors. The Rasch model (which does not incorporate any information about the reported difficulty) performs more similar to the joint models than to the covariate models, suggesting that the latent variables for correctness (θ) and for difficulty perception (h) do not have a strong enough correlation to influence one another. We discuss this further in Sect. 4.2.

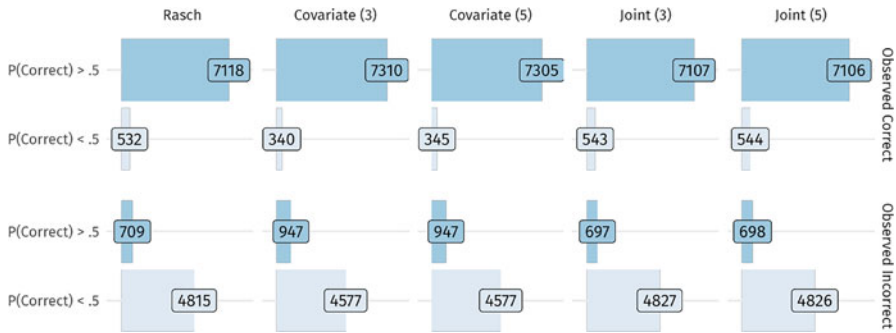


Fig. 3 Predictive performance of each of the five models

Table 3 Estimated out-of-sample prediction error according to WAIC as computed by the `brms` package (Bürkner, 2017). The WAIC for the Joint Models is computed using only ‘correct’ as a response to remain comparable to the other models

	WAIC (y)	\hat{SE} (WAIC)
Covariate (5)	6847.9	123.5
Covariate (3)	6855.4	123.3
Rasch	6873.4	123.6
Joint (3)	6884.3	123.8
Joint (5)	6889.7	123.9

Out-of-sample Predictive Performance The estimated out-of-sample predictive performance is shown in Table 3. The WAIC for the joint response models (“Joint (3)” and “Joint (5)”) have been adjusted to use only the log pointwise predictive density for correctness (e.g., Y_{ij}) so that they remain comparable to the other models and is shown in the first column (labeled $WAIC(y)$). We see that the covariate models perform the best according to WAIC (since they have the lowest values), the joint models perform the worst according to WAIC, and the Rasch model is somewhere in the middle. However, all of the WAIC values are well within a single standard error of one another (as seen in column 2, $\hat{SE}(WAIC)$), suggesting that these differences may be due to sampling error and not true differences in model performance.

4.2 Parameter Estimates

Predictive performance is only one aspect of model evaluation, and while the joint models appear to have more balanced error rates, there is not a clear best-fitting model in terms of prediction error. Here, we compare our person parameter estimates (θ and h) among each of the models, and also investigate the population-level covariates (β 's).

Population-Level Covariates Figure 4 displays the estimated coefficients with 95% posterior intervals. We observe that easier categories tend to have the expected result (higher probability of a correct response), but that items that are rated as ‘Difficult’ or ‘Very Difficult’ (or ‘hard’ in the Covariate (3) model) are expected to have a higher probability of a correct response than items that are rated as ‘medium’. These results suggest that examiners may be over-rating the items that are labeled as ‘very difficult’ or under-rating items labeled as ‘medium’.

Latent Variables Describing Participants The θ estimates for each participant in the Black Box study under each model are shown in Fig. 5, along with 95% posterior intervals. Perhaps unsurprisingly given the structure of the models, each model results in similar estimates for each examiner, and a similar amount of uncertainty in the estimate. There are a couple of examiners who receive higher θ estimates in the joint models as compared to the covariate and Rasch model, which suggests that extreme differences in reported difficulty may lead to different proficiency estimates, but overall these differences are not significant.

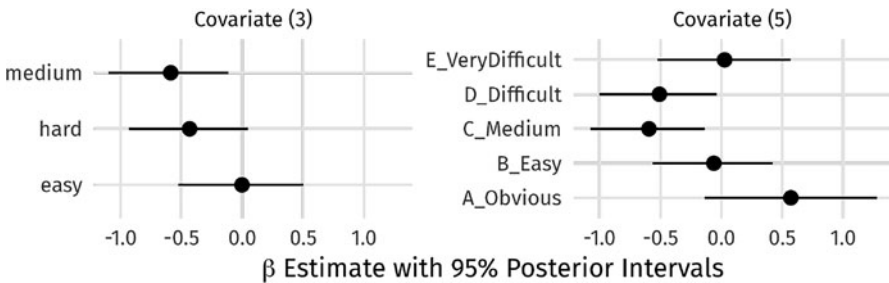


Fig. 4 β Estimates for the covariate models along with 95% posterior intervals

θ Estimates with 95% Posterior Intervals

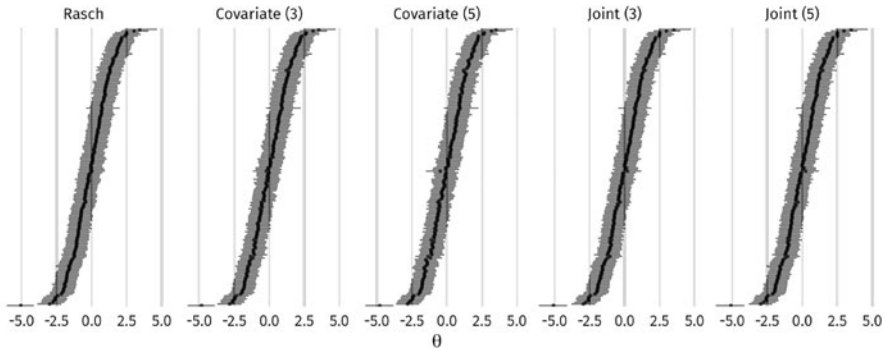


Fig. 5 Proficiency estimates for each model along with 95% posterior intervals. We obtain similar estimates for proficiency in every model

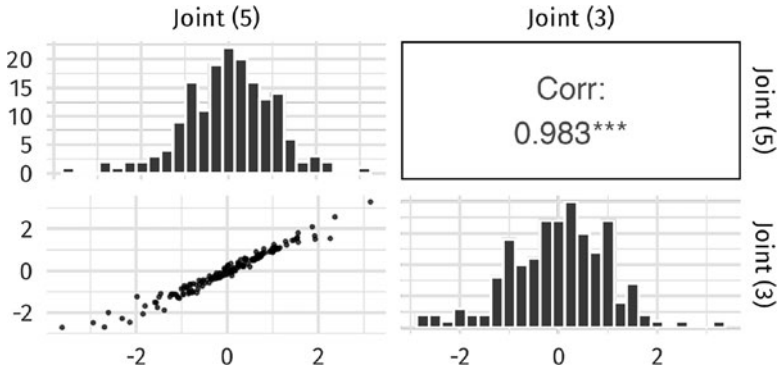


Fig. 6 Comparison of h estimates among the two joint models

Table 4 Correlation between the latent variables in the joint models

Model	Cor(θ, h)	Cor(b, f)
Joint (3)	-0.24 [-0.10, -0.39]	0.48 [0.40, 0.55]
Joint (5)	-0.27 [-0.12, -0.42]	0.47 [0.40, 0.54]

A comparison of the h estimates between the two joint models is shown in Fig. 6. The histograms on the diagonal show the overall distribution of h estimates in the 5-category (left) and 3-category (right) models, and show that the 5-category model may have slight shrinkage of the estimates compared to the 3-category model. We see the same trend in the scatterplot (lower left plot), where the h estimates from the Joint (3) model are displayed on the y -axis and the Joint (5) model is on the x -axis. However, the differences between the two models are quite small and the correlation between the estimates remains extremely high (0.983).

Correlation Between Latent Variables Finally, we investigate the correlation between θ and h and between b and f in the joint models. The correlation estimates, with 95% posterior intervals, are shown in Table 4. The estimates are consistent across the two models, suggesting that the transformed reported difficulty does not lead to substantial changes in overall trends. We also find that there is a stronger correlation between the latent variables describing the items (b and f) than the latent variables describing the participants (θ and h). Items on which participants performed worse (i.e. high b) also tended to be rated as more difficult (i.e. high f). On the other hand, participants who performed better (i.e. high proficiency) tended to rate items as less difficult on average (i.e. lower h), but this relationship is not as strong.

5 Discussion

In this paper, we discussed two broad approaches for incorporating covarying responses within an IRT framework. The first approach involves including the covarying response as a predictor at the same $i \times j$ level as the primary response of interest (e.g. “correctness” in a usual IRT model). The second approach involves treating the outcome as a multivariate response. We used separate latent variables to model each response, but the joint responses could also depend on shared latent variables (see Luby et al. (2020) for one example). We applied each of these models to the FBI ‘Black Box’ study, which measured the performance of forensic fingerprint examiners. In this setting, our primary outcome of interest was whether or not the fingerprint examiner came to the correct source conclusion, and our secondary outcome was the examiner’s reported difficulty of each comparison task.

In terms of predictive performance, including reported difficulty as a population-level covariate offers no improvement to the model. We also find very few differences in conclusions between the models using the original reported difficulty scale and the models using the transformed difficulty scale. The joint models estimated nonzero reporting tendencies for participants (h) and items (f), suggesting that there are detectable differences in reporting behaviors. We prefer the joint modeling approach as obtaining latent variable estimates for reported difficulty provides useful information about variability in reporting behavior.

A limitation with this analysis that we did not address is in scoring the data. While the FBI provided information on whether each pair of fingerprints came from the same source or different sources, there is no keyed correct response. This is especially important when considering “inconclusive” responses. In this analysis, inconclusive responses were treated as incorrect (since the participant was not able to come to the correct conclusion), but in reality we may expect an inconclusive result for very low-quality images. The treatment of inconclusive determinations in determining error rates is an ongoing discussion in the forensic science literature, particularly in pattern evidence disciplines such as fingerprints (Koehler, 2007; Luby, 2019; Dror & Scurich, 2020), palm prints (Eldridge et al., 2021), and firearms (Hofmann et al., 2020).

However, any scoring of the inconclusive responses necessarily results in a loss of information. As an alternative to scoring inconclusive results as correct or incorrect, it is also possible to use a consensus-based approach or ‘IRT without an answer key’ (Batchelder & Romney, 1988; Luby et al., 2020). This is another potential avenue of future research, since how the responses are scored has a significant impact on the estimated proficiency and difficulty estimates, and we may expect a similar effect for estimating secondary latent variables.

A further challenge in modeling joint responses in general, particularly in application areas in which it is infeasible to undergo substantial validation testing, is that using the raw secondary response may lead to overfitting. As noted in Sect. 3, reported difficulty was not consistent for examiner \times item pairs in a follow-up repeatability study using the same participants and item pool (Ulery et al., 2012). In

future studies, providing an opportunity for participants to calibrate their secondary responses may improve inferences.

In the future, we would also like to investigate whether there is a differential use of subjective reporting scales in forensic science. Allowing the category thresholds in Eq. 6 to vary across participants would be one way of accounting for differential use of the reporting scales.

Even with the challenges outlined above, it is important to study these secondary responses and better understand the variability among participants and items. While it is not standard to collect a reported difficulty on a five-point scale in forensic casework, it is common for examiners to testify in court regarding their conclusions. If there is substantial variability in how examiners perceive the difficulty of fingerprint comparisons, this may lead to variability in testimony that is provided to judges and juries. Collecting and modeling covarying information (reported difficulty or otherwise) could provide additional insight into the differences in perception and decision-making than responses alone.

References

- AAAS. (2017). Forensic science assessments: A quality and gap analysis - latent fingerprint examination. Tech. rep., (prepared by William Thompson, John Black, Anil Jain, and Joseph Kadane)
- Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika*, 53(1), 71–92.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28.
- Bürkner, P. C. (2019). Bayesian item response modeling in R with brms and Stan. Preprint, arXiv:190509501.
- De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software, Code Snippets*, 48(1), 1–28. <https://doi.org/10.18637/jss.v048.c01>, <https://www.jstatsoft.org/v048/c01>
- De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.
- Dror, I. E., & Scurich, N. (2020). (Mis) use of scientific measurements in forensic science. *Forensic Science International: Synergy*, 2, 333–338.
- Eldridge, H., De Donno, M., & Champod, C. (2021). Testing the accuracy and reliability of palmar friction ridge comparisons—a black box study. *Forensic Science International*, 318, 110457.
- Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement*, 31(6), 525–543. <https://doi.org/10.1177/0146621606295197>
- Fischer, G. H., & Molenaar, I. W. (2012). *Rasch models: Foundations, recent developments, and applications*. New York: Springer Science & Business Media.
- Friction Ridge Subcommittee of the Organization of Scientific Area Committees for Forensic Science. (2017). Guideline for the articulation of the decision-making process leading to an expert opinion of source identification in friction ridge examinations. Online; accessed September 15, 2021.

- Friction Ridge Subcommittee of the Organization of Scientific Area Committees for Forensic Science. (2019). Friction ridge process map (current practice). Online; accessed September 15, 2021.
- Hofmann, H., Carriquiry, A., & Vanderplas, S. (2020). Treatment of inconclusives in the AFTE range of conclusions. *Law, Probability and Risk*, 19(3–4), 317–364.
- Holland, P. W., & Wainer, H. (2012). *Differential item functioning*. Routledge.
- Jeon, M., De Boeck, P., & van der Linden, W. (2017). Modeling answer change behavior: An application of a generalized item response tree model. *Journal of Educational and Behavioral Statistics*, 42(4), 467–490.
- Koehler, J. J. (2007). Fingerprint error rates and proficiency tests: What they are and why they matter. *Hastings LJ*, 59, 1077.
- Luby, A. (2019). Decision making in forensic identification tasks. In S. Tyner & H. Hofmann (Eds.), *Open forensic science in R* (Chap. 13). rOpenSci, US.
- Luby, A., Mazumder, A., & Junker, B. (2020). Psychometric analysis of forensic examiner behavior. *Behaviormetrika*, 47, 355–384.
- Luby, A., Mazumder, A., & Junker, B. (2021). Psychometrics for forensic fingerprint comparisons. In *Quantitative psychology* (pp. 385–397). Springer.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174.
- R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. <http://www.R-project.org/>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.
- Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, 30(4), 298–321.
- Stan Development Team. (2018a). RStan: The R interface to Stan. r package version 2.18.2. <http://mc-stan.org/>
- Stan Development Team. (2018b). Stan modeling language users guide and reference manual. <http://mc-stan.org>
- Thissen, D. (1983). 9 - timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 179–203). San Diego: Academic.
- Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2011). Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences*, 108(19), 7733–7738.
- Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2012). Repeatability and reproducibility of decisions by latent fingerprint examiners. *PloS One*, 7(3), e32800.
- 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., Klein Entink, R. H., & Fox, J. P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, 34(5), 327–347.
- 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.
- Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec), 3571–3594.

Exploring the Utility of Nonfunctional Distractors



Merve Sarac and Richard A. Feinberg

Abstract Functional distractors (the incorrect options in a multiple-choice question) should draw attention from those test-takers who lack sufficient ability or knowledge to respond correctly. Unfortunately, previous research on distractors has demonstrated the unsettling reality that this rarely occurs in practice leading to recommendations for creating items with fewer incorrect alternatives. The purpose of the present study was to explore if these nonfunctional distractors (NFDs) may still yield value in detecting unusual examinee behavior. Using empirical data from a high-stakes licensure examination, examinees who selected an excessive number of NFDs were flagged and analyzed with respect to their response times and overall performance. Results indicated that these flagged examinees were also of extremely low ability, selected NFDs consistently across item sequence, and were homogenous in their pacing strategies - spending a similar amount of time when choosing a nonfunctional or functional distractor. Implications for relevant policy decisions, mitigation strategies, operational applications, and test security considerations are discussed.

Keywords Distractor quality · Aberrant responses · Response time · Multiple-choice test

Multiple-choice questions (MCQs) have been ubiquitous in standardized testing over time and across various disciplines due to their cost-effective development, intuitive presentation, and objective scoring (e.g., Gierl et al., 2017). Given modern advances in automated item generation (e.g., von Davier, 2018; Prasetyo et al., 2020) and distractor generation (e.g., Susanti et al., 2018; Maurya & Desarkar, 2020),

M. Sarac (✉)

University of Wisconsin-Madison, Department of Educational Psychology, Madison, WI, USA
e-mail: sarac@wisc.edu

R. A. Feinberg

National Board of Medical Examiners, Philadelphia, PA, USA
e-mail: rfeinberg@nbme.org

MCQs will likely continue to be widely used in the foreseeable future. MCQs are generally composed of a stem or prompt followed by a series of discrete response options, one of which being the (most) correct answer. While the literature on constructing quality MCQs is vast, historically, the focus has been on developing the stem and correct option, though studies on the incorrect options or distractors have been gaining attention (Thissen et al., 1989; Gierl et al., 2017).

An effective or functional distractor should draw attention from those test-takers who lack sufficient ability or knowledge to respond correctly. Previous research on distractors has demonstrated the unsettling reality that in many contexts they tend to be statistically nonfunctional (e.g., not disproportionately attracting examinees who lack ability or knowledge), prompting practitioners to question their necessity (Delgado & Prieto, 1998; Haladyna et al., 2002; Raymond et al., 2019; Rodriguez, 2005; Rogausch et al., 2010). For instance, a traditional MCQ presentation would include at least four options with three distractors to mitigate correct guessing, thereby making exam scores more valid and reliable (Haladyna & Downing, 1993; Rodriguez, 2005). However, nonfunctional distractors (NFDs) are so implausible that even low ability test-wise examinees can easily eliminate them to increase their probability of correct guessing (Delgado & Prieto, 1998; Haladyna & Downing, 1993; Rodriguez, 2005). Further, NFDs introduce other costs and risks related to the item development efforts in creating the distractors, allotting sufficient testing time for examinees to read all response options, and potentially exacerbating negative perceptions about a test's integrity with incredulous response options. For these reasons, researchers have investigated the optimal number of distractors for MCQs (Delgado & Prieto, 1998; Haladyna et al., 2002; Raymond et al., 2019; Rogausch et al., 2010) with Rodriguez (2005) concluding after a meta-analysis across 80 years of research that three-response options are ideal.

Despite NFD prevalence, distractor analyses are routinely carried out in operational testing programs to improve item quality (Haladyna, 2016). It is well documented that the item information function under an item response theory (IRT) framework is affected by distractor characteristics such as difficulty and discrimination (Haladyna & Downing, 1988; Revuelta, 2004). Additionally, items with distractors of similar difficulty are more informative about an examinees' ability than those with varying difficulty levels. Beyond improving test development practices, modeling the propensity to choose distractors has been shown to increase measurement precision (Levine & Drasgow, 1983; Thissen & Steinberg, 1984). For example, several IRT models for distractors were developed in the literature (Bock, 1972; Briggs et al., 2006; Haberman & Lee, 2017; Samejima, 1979; Suh & Bolt, 2010; Thissen et al., 1989). Additionally, to extract diagnostic information on examinee learning deficits or misconceptions, cognitive diagnostic models (CDMs) have been proposed to systematically develop and analyze distractors (de la Torre, 2009). Distractor analyses also improve the detection of unusual response similarity in the context of test security (e.g., Haberman & Lee, 2017; Wollack, 1997). Several research studies use information from distractors in addition to the correct options to identify aberrant patterns of matching responses between examinees (e.g., ω

(Wollack, 1997), M4 (Maynes, 2014), and Generalized Binomial Test (GBT: van der Linden & Sotaridona, 2006).

Though previous research recommends replacing or eliminating NFD to implement MCQs with fewer response options, distractor analyses are utilized in various ways by operational testing programs to optimize score validity. The purpose of the present study is to extend this distractor literature by focusing specifically on the potential value of NFDs. Particularly, we highlight the extent to which excessive NFD selection may serve as a mechanism to detect unusual response behavior. Thus, illustrating how NFDs are, counter-intuitively, worth retaining for forensic analyses.

1 Methods

Item response data was obtained from a licensure examination in which the primary purpose is a pass/fail classification. As a high-stakes assessment, prospective examinees were well-informed of the non-trivial financial cost (over \$500) to register as well as the consequences for a failing attempt. Thus, examinees should be highly motivated to only test when they are ready to perform to the best of their ability. Data included responses from 61,088 examinees, each completing a form composed of 280 MCQs in which most (97%) had either four, five, or six response options. MCQs were randomly administered across seven 40-item blocks with 60 minutes allotted to complete each block.

First, distractors across all items were categorized as either functional or non-functional. Though several definitions have been proposed in the literature, a simple and commonly used approach was employed to classify nonfunctional distractors as any incorrect option selected by less than 5% of the group (Wakefield, 1958). Using this NFD definition, the total number of NFDs selected by each examinee and their corresponding rank relative to the entire group was calculated. Zipf's law (1949) was used to detect outliers among examinees who selected an excessive number of NFDs. Zipf's law states that the relationship between the log of the frequency of an event and its rank among other events is linear (i.e., a linear relationship between the number of occurrences of NFD and its rank ordering), a relationship that holds reasonably well in many situations (Wainer, 2000; Aitchison et al., 2016). The top 1% of all examinees who selected NFDs abundantly more than expected from the linear relationship between the NFD log frequency and its rank were flagged.

The remaining analyses focus on the flagged examinees who engaged in excessive NFD selection. First, the relationship between NFD selection and examinees' response times was investigated as previous research has demonstrated connections between response time and motivation (Wise & Kuhfeld, 2020) and speededness (Feinberg et al., 2021), both of which may be potential causal factors for NFD selection. Average standardized response times relative to an item's time-intensity were calculated for all examinees for only the responses in which they selected an NFD. Flagged examinees were then further subdivided into three groups: fast (their average standardized NFD response time in seconds was greater than 0.5), slow

(their average standardized NFD response time was less than -0.5), or typical (their average standardized NFD response time was between -0.5 and 0.5). Next, overall test performance (i.e., total scores) was compared against NFD selection to detect conditional differences across ability levels, potentially informing a stopping-rule criterion. Lastly, the relationship between NFD selection and item characteristics, such as item difficulty and stem length (in word counts), was explored to explain why some items received higher rates of NFD selection.

2 Results

Following the definition of a distractor as nonfunctional if selected by less than 5% of examinees, we described the number of functional distractors by MCQ option set (Table 1). For instance, in the 5-option set, only 5% of the items had all functional distractors, 82% between 1 and 3 functional distractors, and 14% had no functional distractor (i.e., all were NFDs). These results aligned with the literature that many incorrect responses were not distracting well. Items with no functional distractors can be considered extremely easy - typically in which the proportion of the group responding correctly (p value) is greater than 90% (Gierl et al., 2017). Easy items tend to commonly appear in licensure and certification tests due to content coverage requirements.

Figure 1 shows the log-frequency of NFDs selected by their rank, revealing an exponential increase as the percentile approached 100. Highlighted in red were examinees at or above the 99th percentile for how often they selected an NFD. A total of 603 examinees were flagged as excessively selecting NFDs more than would be expected – on at least 64 of the 280 MCQs.

When selecting NFDs, these flagged examinees were categorized as either fast, slow, or typical relative to the item's standardized response time (Fig. 2). Examinees highlighted in green were very slow but ultimately selected an NFD. Perhaps they lacked sufficient knowledge yet engaged for a lengthy amount of time before providing a misinformed response. In comparison, examinees in red were those who picked an NFD quickly. One explanation could be that they incidentally selected NFDs from a rapid random guess due to speededness.

Table 1 Percent of distractors that are functional for MCQs with 4, 5, or 6 response options

# of Distractors	4 Option Sets (n = 697)	5 Option Sets (n = 5426)	6 Option Sets (n = 1023)
5	–	–	2%
4	–	5%	8%
3	13%	17%	19%
2	40%	31%	29%
1	33%	34%	30%
0	13%	14%	12%

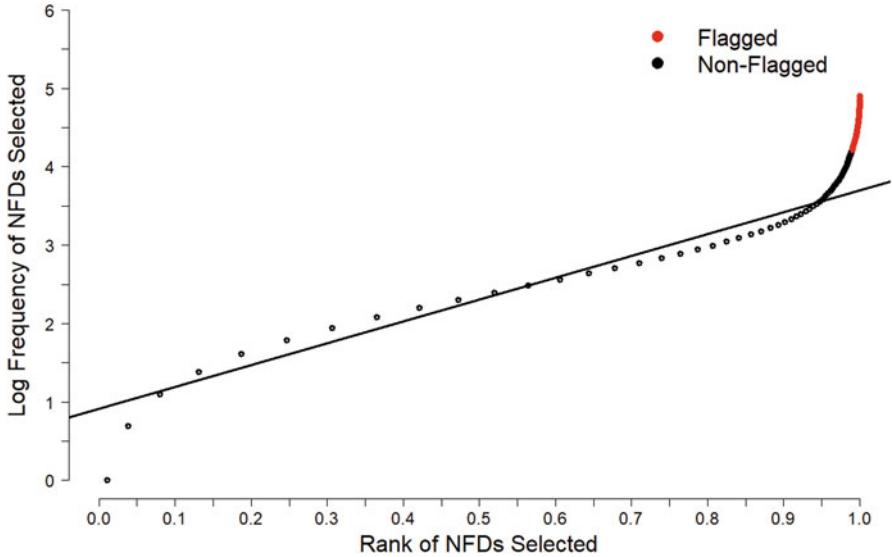


Fig. 1 Log frequency of NFDs by their rank

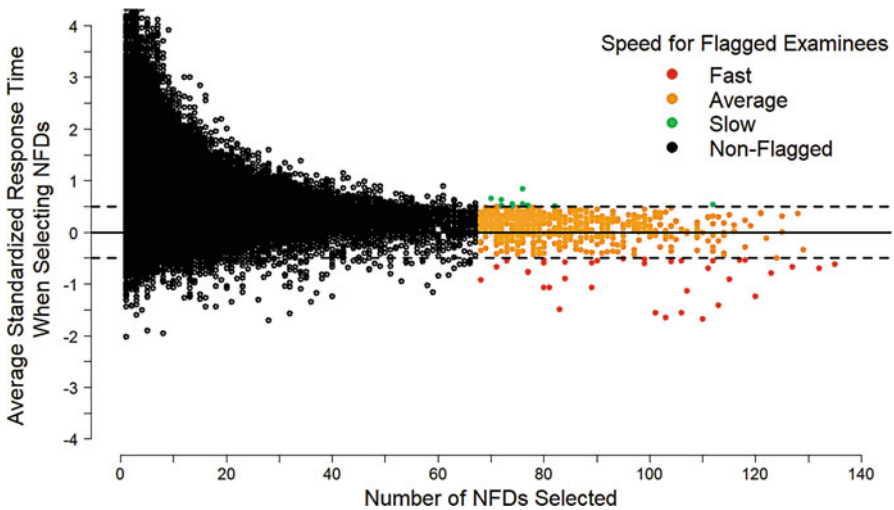


Fig. 2 Average standardized response time when selecting NFDs by number of NFDs selected

However, this explanation seems unlikely given the results presented in Fig. 3; fast responder flagged examinees also tended to complete the test with ample remaining time. Not shown here, but the percentage of examinees by flagged group choosing an NFD was stable across item encounter sequence - NFD selection was not more prevalent at the end of each section. Thus, the results suggest NFD

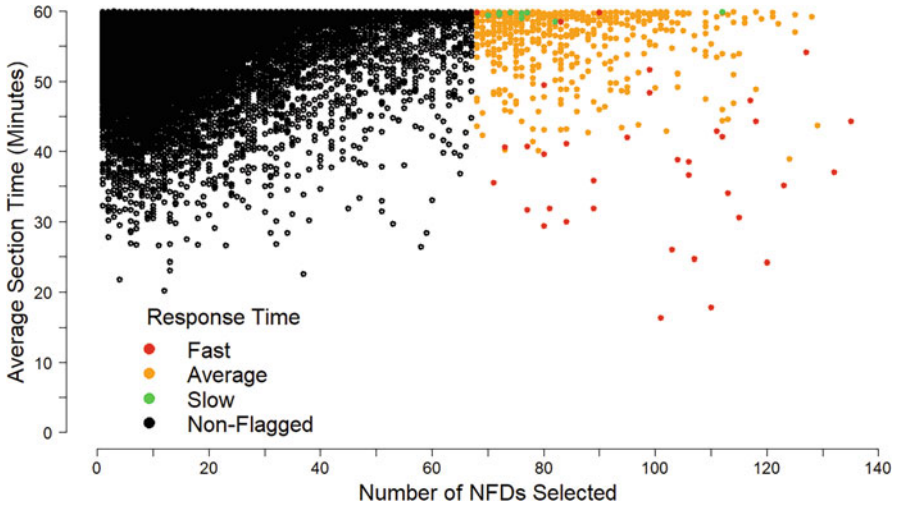


Fig. 3 Average section response time by number of NFDs selected

selection was not influenced by speededness (Bolt et al., 2002; Harik et al., 2020; Feinberg et al., 2021).

The relationship between total test score and the number of NFDs selected (Fig. 4) indicated that flagged examinees were also those with the lowest overall performance, well below the passing score threshold. All those flagged failed the test, in addition to any examinees with more than 45 NFD selections. This finding also suggested that examinees who selected a high number of NFDs might have been the least prepared for the test. If so, then for low ability examinees NFDs were indeed functional as they were helping to distinguish uninformed responders.

Figure 5 illustrates the relationship among total test score by the response time difference between NFD and functional distractors. The plot indicates that the time examinees spent choosing an incorrect option that was nonfunctional vs. functional varied more for examinees of higher ability.¹ Flagged examinees, who also happened to be of the lowest ability, had a relatively small variance in pacing (i.e., their pacing is similar for when they selected a functional or nonfunctional distractor). Thus, slow flagged examinees spent a long time on all incorrect responses, and fast flagged examinees spent very little time on all incorrect responses, regardless of whether they chose an NFD.

Lastly, Table 2 shows how the frequency of NFD selection for flagged examinees related to item difficulty (b-parameter in the Rasch model) and stem length in word counts (mean-centered at 117.4 words). Due to non-normality of NFD frequency, bootstrap confidence intervals were provided for the linear regression estimates.

¹ Note that this relatively higher variance may also be an artifact of increased standard error due to selecting fewer incorrect responses in general.

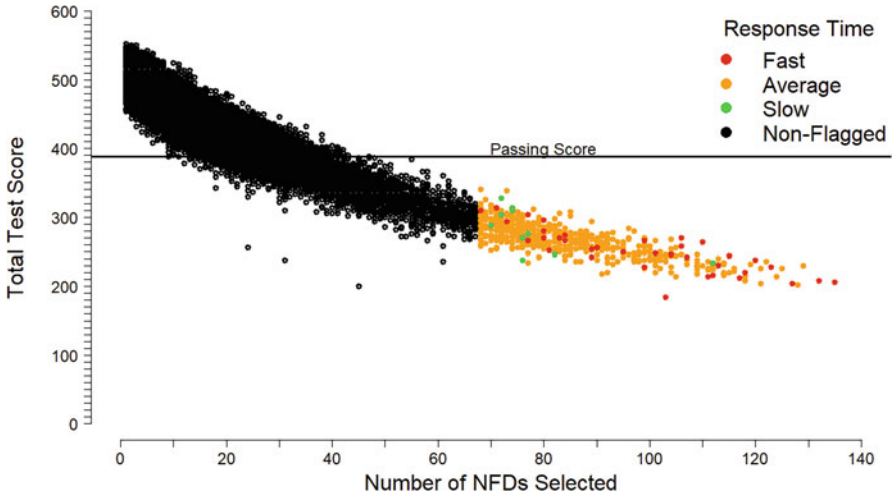


Fig. 4 Total test score by the number of NFDs selected

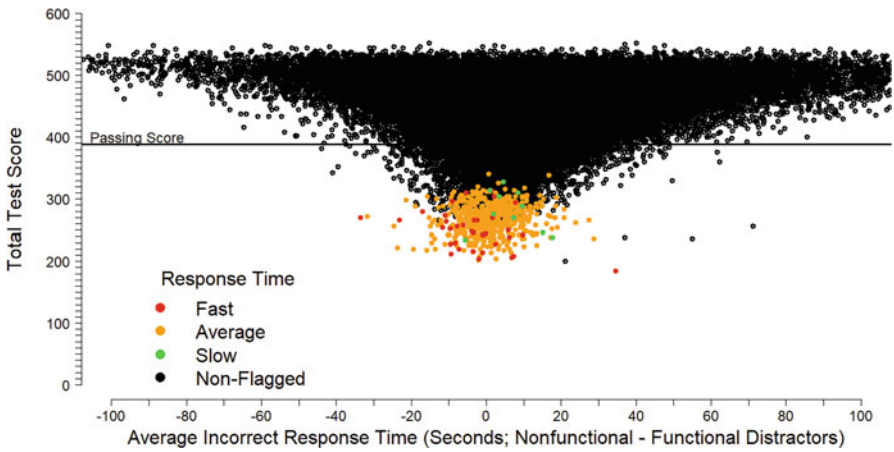


Fig. 5 Total test score by nonfunctional and functional distractor response time difference

Table 2 Summary of the linear regression of item NFD frequency

	Estimate	95% CI
Constant	10.61*	[10.02, 11.22]
Item difficulty (b-parameter)	-4.28*	[-5.36, -3.37]
Stem length (in word counts)	-0.03*	[-0.05, -0.02]

Note: * $p < .001$. Stem length is mean-centered ($M = 117.4$)

Item difficulty was negatively associated with the number of NFDs selected after controlling for stem length ($\beta = -4.28$). As item difficulty (b-parameter) decreased, the number of NFDs selected increased (i.e., easier items tended to have

more NFDs selected). Item stem length was also negatively associated with the number of NFDs selected, controlling for item difficulty ($\beta = -0.03$). As stem length decreased, the number of NFDs selected increased.

3 Discussion

Though NFDs can be prevalent among MCQ response options, they may offer value for posttest forensic analysis and insight into examinee behavior. Results from this study suggest that flagged examinees who excessively selected NFDs were also of extremely low ability, selected NFDs consistently across item sequence, and were homogenous in their pacing strategies - spending a similar amount of time when choosing an NFD or functional distractor. Additionally, many of the flagged examinees had ample time remaining, which could have been used to consider a more attractive response option, perhaps even the correct answer.

Response time investigations revealed that among those flagged with extreme NFD selection, slow responders used most if not all the time available while fast responders had plenty of time remaining. Causal factors to explain these different behavior patterns, such as incorrect content preknowledge, malicious test-taking behavior (e.g., item harvesting) or purposeful failure to perpetuate retest opportunities and gain exposure to secure test material, would be important to investigate in future studies. If the methods described in this paper were used operationally, practitioners may want to further subdivide the flagged examinees, if possible, by other relevant characteristics (e.g., school, testing location, or native language), which might provide additional insight into explanatory factors. Researchers could also investigate the retest performance (or passing rates) for the flagged examinees to distinguish between low ability and aberrant behavior. Further exploration could also include analyses on the match on NFD selections among those flagged, which could relate to the (incorrect) content preknowledge or collusion.

This study used only one frequency-based definition of NFDs as recommended by Wakefield (1958). Results may differ across more sophisticated or conservative approaches that incorporate other ancillary information such as item discrimination, biserial correlations, or an average total score of those selecting the distractor (e.g., Haladyna & Downing, 1988). Additionally, licensure and credentialing tests similar to the one used in this study tend to have easier items for a more homogenous testing population, leading to relatively more NFDs (based on a response-frequency NFD definition) than one might expect with a heterogeneous achievement testing population. Thus, results from this study should be interpreted with caution in how they may generalize to other contexts.

As automated item generation (AIG) models and techniques, mostly implemented using natural language processing (NLP: Prasetyo et al., 2020), are receiving growing attention in various disciplines of standardized testing, algorithms allowing for automated generation of distractors are becoming more practical. Several

approaches in the AIG literature have been focused on generating quality distractors that are in context with items but semantically dissimilar to the correct options (Maurya & Desarkar, 2020). Thus, as the creation of distractors evolves over time, additional research will be needed on the mechanisms to evaluate their efficacy and appropriateness for different purposes.

Findings from this study support the utility of NFDs for identifying aberrant examinee behavior. High NFD selection could occur for various reasons, such as low motivation, population knowledge deficiencies, incorrect preknowledge, purposeful poor performance, or malicious content harvesting. Though the intent of the examinees is challenging to ascertain, results from this study highlight that NFDs could be used to inform test security procedures, particularly for licensure and credentialing exams. For instance, a within-day stopping rule could be implemented without pass/fail implications (e.g., after an examinee selects a predetermined number of NFDs). Further, limiting the number of repeating attempts or the time between attempts may also be warranted. Efforts to restrict content exposure may also benefit non-malicious yet unprepared examinees, who would be best served by mitigating their risk for academic consequences due to additional failing attempts. Thus, using NFDs as a mechanism to identify aberrant behavior could be mutually beneficial to both testing programs and examinees by maintaining the health of item banks through decreased exposure rates, preserving the unidimensionality of a test by mitigating construct-irrelevant factors, and preventing examinees from retesting who need substantial remediation.

References

- Aitchison, L., Corradi, N., & Latham, P.E. (2016). Zipf's law arises naturally when there are underlying, unobserved variables. *PLoS Computational Biology*, *12*(12), e1005110. <https://doi.org/10.1371/journal.pcbi.1005110>
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51. <https://doi.org/10.1007/bf02291411>
- 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. <https://doi.org/10.1111/j.1745-3984.2002.tb01146.x>
- Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, *11*, 33–63. https://doi.org/10.1207/s15326977ea1101_2
- de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple choice options. *Applied Psychological Measurement*, *33*, 163–183. <https://doi.org/10.1177/0146621608320523>
- Delgado, A., & Prieto, G. (1998). Further evidence favoring three-option items in multiple-choice tests. *European Journal of Psychological Assessment*, *14*, 197–201. <https://doi.org/10.1027/1015-5759.14.3.197>
- Feinberg, R. A., Jurich, D. P., & Wise, S. L. (2021). Reconceptualizing rapid responses as a speededness indicator in high-stakes assessments. *Applied Measurement in Education*. Advance online publication. <https://doi.org/10.1080/08957347.2021.1987904>

- Gierl, M. J., Balut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research, 87*, 1082–1116.
- Haberman, S. J., & Lee, Y. -H. (2017). A statistical procedure for testing unusually frequent exactly matching responses and nearly matching responses (Research Report No. RR-17-23). Educational Testing Service. <https://doi.org/10.1002/ets2.12150>.
- Haladyna, T. M. (2016). Item analysis for selected-response test items. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 392–409). Routledge.
- Haladyna, T. M., & Downing, S. M. (1988, April). *Functional distractors: Implications for test-item writing and test design* [Paper presentation]. American Educational Research Association.
- Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice item? *Educational and Psychological Measurement, 53*, 999–1010.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item writing guidelines for classroom assessment. *Applied Measurement in Education, 15*, 309–333. https://doi.org/10.1207/S15324818AME1503_5
- Harik, P., Feinberg, R. A., & Clauser, B. E. (2020). How examinees use time: Examples from a medical licensing examination. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 73–89). Routledge.
- Levine, M. V., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement, 43*, 675–685. <https://doi.org/10.1177/001316448304300301>
- Maurya, K. K., & Desarkar, M. S. (2020, October). Learning to distract: A hierarchical multi-decoder network for automated generation of long distractors for multiple-choice questions for reading comprehension. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 1115–1124).
- Maynes, D. D. (2014). Detection of non-independent test taking by similarity analysis. In N. M. Kingston & A. K. Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 53–82). Routledge.
- Prasetyo, S. E., Adji, T. B., & Hidayah, I. (2020, September). Automated item generation: Model and development technique. In *The 7th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)* (pp. 64–69).
- Raymond, M. R., Stevens, C., & Bucak, D. S. (2019). The optimal number of options for multiple-choice questions on high-stakes tests: Application of a revised index for detecting nonfunctional distractors. *Advances in Health Sciences Education, 24*(1), 141–150.
- Revuelta, J. (2004). Analysis of distractor difficulty in multiple-choice items. *Psychometrika, 69*(2), 217–234.
- Rodriguez, M. C. (2005). Three options are optimal for multiple-choice test items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3–13.
- Rogausch, A., Hofer, R., & Krebs, R. (2010). Rarely selected distractors in high stakes medical multiple-choice examinations and their recognition by item authors: A simulation and survey. *BMC Medical Education, 10*, 85. <https://doi.org/10.1186/1472-6920-10-85>
- Samejima, F. (1979). *A new family of models for the multiple-choice item* (Office of Naval Research Report Nos. 79-4, N00014-77-C-0360). University of Tennessee, Department of Psychology.
- Suh, Y., & Bolt, D. M. (2010). Nested logit models for multiple-choice item response data. *Psychometrika, 75*, 454–473. <https://doi.org/10.1007/s11336-010-9163-7>
- Susanti, Y., Tokunaga, T., Nishikawa, H., & Obari, H. (2018). Automatic distractor generation for multiple-choice English vocabulary questions. *Research and Practice in Technology Enhanced Learning, 13*(1), 1–16.
- Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*, 501–519. <https://doi.org/10.1007/bf02302588>
- Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement, 26*, 161–176.

- van der Linden, W. J., & Sotaridona, L. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics, 31*(3), 283–304.
- von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika, 83*(4), 847–857.
- Wainer, H. (2000). Rescuing computerized testing by breaking Zipf's law. *Journal of Educational and Behavioral Statistics, 25*(2), 203–224.
- Wakefield, J. A. (1958). Does the fifth choice strengthen a test item? *Public Personnel Review, 19*, 44–48.
- Wise, S. L., & Kuhfeld, M. R. (2020). A cessation of measurement: Identifying test taker disengagement using response time. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 150–164). Routledge.
- Wollack, J. A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement, 21*, 307–320. <https://doi.org/10.1177/01466216970214002>
- Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley.

Continuation Ratio Model for Polytomous Items Under Complex Sampling Design



Diego Carrasco , David Torres Iribarra, and Jorge González 

Abstract The use of polytomous items as part of background or context questionnaires and complex sampling designs are two features common in international large-scale assessments (ILSA). Popular choices to model polytomous items within ILSA include the partial credit model, the graded response model, and confirmatory factor analysis. However, an absent model in ILSA studies is the continuation ratio model. The continuation ratio model is a flexible alternative and a very extendable response model applicable in different situations. Although existing software can fit this model, not all these tools can incorporate complex sampling design features present in ILSA studies. This study aims to illustrate a method to fit a continuation ratio model including complex sampling design information, thus expanding the modelling tools available for secondary users of large-scale assessment studies.

Keywords Continuation Ratio Model · Polytomous items · Item response theory · Bullying

D. Carrasco (✉)

Centro de Medición MIDE UC, Pontificia Universidad Católica de Chile, Santiago, Chile
e-mail: dacarras@uc.cl

D. T. Iribarra

Escuela de Psicología, Pontificia Universidad Católica de Chile, Santiago, Chile

Millennium Nucleus on Intergenerational Mobility: From Modelling to Policy (MOVI), Vitacura, Chile

e-mail: davidtorres@uc.cl

J. González

Departamento de Estadística, Pontificia Universidad Católica de Chile, Santiago, Chile

Millennium Nucleus on Intergenerational Mobility: From Modelling to Policy (MOVI), Vitacura, Chile

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

1 Introduction

Most of the items contained in background and context questionnaires in large-scale assessment (ILSA) (Rutkowski et al., 2010) use a polytomous response format with ordered options. Popular response models used to generate scale scores based on the ordered responses from these questionnaires are the partial credit model (PCM, Masters, 1982), the graded response model (GRM, Samejima, 1968) and confirmatory factor analyses (CFA, Jöreskog, 1969) which have been used in TIMSS, PIRLS and ICCS (IEA), TERCE (UNESCO), in LSCE (UNICEF) and in TALIS (OECD); (for a summary see Table 1). However, an absent model in ILSA studies is the continuation ratio response model (CRM, Tutz, 2016). This item response model, in its constrained form, satisfies some of the Rasch model properties, such as stochastic ordering (Hemker et al., 2001), parameter separability (Tutz, 1990), and monotonicity (Van Der Ark, 2001). The continuation ratio models can include different random terms within the generalized latent variable framework, the typical continuous latent variable, and mixture random terms (Masyn, 2003), resulting in a very extendable and flexible response model applicable to different situations. This model is also called the sequential model (Tutz, 2016), and it can be seen as a special case of an IRTree model (Jeon & De Boeck, 2016). We believe that the sparse adoption of the CRM is related to its lack of availability in off-the-shelf software. Although there are tools that can fit this model, including, for instance, R libraries such as *mirt* (Chalmers, 2012) and *lme4* (Bates et al., 2015; De Boeck et al., 2011), the program *gllamm* in STATA (Rabe-Hesketh et al., 2004), and generalized latent variable software such as LatenGold (Vermunt & Magidson, 2016), not all these tools can handle complex sampling design features as those usually seen in ILSA studies. In this paper, we illustrate a method to fit a continuation ratio model, based on data expansion techniques (Gochyyev, 2015). Our aim is to expand the modelling tools available for secondary users of large-scale assessment studies.

The paper is organized as follows: In section “[Polytomous Scales and Item Response Theory Models](#)” we describe a common background multi-item bullying scale with polytomous ordered responses, for which the continuation ratio response model is applicable. Item response theory models that have been used in ILSA studies, and the Continuation Ratio Model are also briefly described in terms of the way the logit link is used in their specification. Section “[Inferences to the Population](#)” describes how the estimates generated with this response model can be used to make inferences about a population. Then, in section “[Methods](#)”, we describe in more detail the data used for illustration purposes, alongside the expansion technique used to fit the continuation ratio response model. In section “[Results](#)” we describe the obtained results. We conclude in section “[Conclusion and Discussion](#)” with discussing some pertinent expansions of this model to its application on bullying scales.

Table 1 Summary of latent variable models used to generate scale scores with context questionnaires

ILSA study	PCM	CFA	GRM
TIMSS 2019	X		
PIRLS 2016	X		
PISA 2018	X		
ICCS 2019	X		
TALIS 2018		X	
TERCE 2013		X	
LSCE 2019			X

Note: *ILSA* international large scale assessment study, *PCM* partial credit model, *CFA* confirmatory factor analysis, *GRM* graded response model, *TIMSS* trends in international mathematics and science study, *PIRLS* progress in international reading literacy study, *PISA* program for international assessment, *ICCS* international civic and citizenship education study, *TALIS* teaching and learning international survey, *TERCE* third regional comparative and explanatory study, *LSCE* life skills and citizenship education study

2 Polytomous Scales and Item Response Theory Models

2.1 An Applied Example of Polytomous Scale

The battery of items “Students’ experience of physical and verbal abuse at school” is a good example of a multi-item scale with polytomous response options present in different ILSA studies. It contains six different items, a frame of reference of the last 3 months, and four ordered response options. This collection of items comes from the International Civic and Citizenship Education Study, commonly abbreviated as ICCS 2016. The chosen bullying scale is an example of a students’ school victimization battery, which collects responses regarding how often students have suffered from different events of bullying at their school. For short, we will call this scale the “bullying scale” throughout this manuscript. Similar instruments are present in other ILSA studies, such as the “Student bullying scale” in TIMSS 2019 (Yin & Fishbein, 2020), and similar versions of this battery are present both in PIRLS 2016, “Student bullying scale” (Martin et al., 2017) and in PISA 2018 (“Exposure to Bullying”; OECD, 2019).

Different item response theory models can be used to handle polytomous responses generated with this type of instruments. For instance, PISA 2018 relied on a generalized partial credit model (OECD, 2020), while both TIMSS 2019 and PIRLS 2016 used a partial credit model for their “Student bullying scale” (Martin et al., 2017; Yin & Fishbein, 2020). Similarly, ICCS 2016 also used a partial credit model to assess the propensity of “Students’ experience of physical and verbal abuse at school” (Schulz et al., 2018).

2.2 Item Response Models

Multiple item response theory models have been developed to deal with polytomous items with ordered response options. Among them, three of the most common models are characterized by the use of different formulations of the logit link (De Boeck & Partchev, 2012): the partial credit model (PCM) is based on the adjacent ratio logit, the graded response model (GRM) on the cumulative ratio logit, and the sequential model on the continuation ratio logit. The adjacent ratio generates logits of the odds for contiguous categories of responses (1 vs 2; 2 vs 3; 3 vs 4); while the cumulative ratio logits are logits of the odds for 1 vs higher categories, 1–2 vs higher, and 1–2–3 vs 4. Finally, the continuation ratio logit generates logits for the odds of the second category and its previous response among the response options, the third category and its previous, and the last category and all its previous. In the present study, we are interested in this last formulation.

2.3 Continuation Ratio Model

Continuation Ratios can be specified in increasing or decreasing order (Gochyyev, 2015). This modelling decision changes the interpretation of the item parameters, and of the generated realizations of the response model.

We consider the continuation ratio logit link is applicable to model the response options of the bullying scale given that the structure of the ordered response options: “none at all” should have occurred before a “once” response; a “2 to 4 times” should have occurred before a “5 or more times”, thus conforming to the assumed sequential process (Agresti, 2019).

If the response options of the chosen “bullying scale” are numerically coded from zero to three, expressing the higher frequency of each event, then “not at all” would be zero, “once” would be 1, “2 to 4 times” would be 2, and “5 times or more” would be 3. Using these numeric codes, we can express the continuation ratio link for each of the response options formally.

Let y_{pi} be the response of student p to item i from the chosen “bullying scale”, and θ_p models the students’ propensity of being bullied, while δ_i models how frequent the event of bullying across the students is, in comparison to its earlier category of response. Formally, we can express the continuation ratio in increasing order in the following way:

$$\log \left(\frac{\Pr(y_{pi} \geq s)}{\Pr(y_{pi} = s - 1)} \right) = \theta_p - \delta_{is}, \quad s = 0, 1, 2, 3 \tag{1}$$

A continuation ratio model specified for response options in increasing order would compare all the categories, including its target category, to its previous category. Thus, if our target category is “once”, the estimated logits express the chances of

being bullied at least once in 3 months, in comparison to not suffering that form of bullying at all.

In the present study, we are interested in the decreasing ordered specification. When using the same numeric codes for each response category and specifying a decreasing order for the continuation ratio link, θ_p models the students' propensity of being bullied, while δ_i models how frequent the event of bullying across the students is, in comparison to all earlier frequency options. Formally, we can write this model specification as follows:

$$\log \left(\frac{\Pr(y_{pi} = s)}{\Pr(y_{pi} < s)} \right) = \theta_p - \delta_{is}, \quad s = 0, 1, 2, 3 \quad (2)$$

This last specification is formally similar to how survival and hazard models use the continuation ratio (Agresti, 2019; Masyn, 2014), where the numerator of the odds is a single category, and the denominator of the odds varies as the category of response of interest is higher.

3 Inferences to the Population

One of the main aims of ILSA is to enable comparison between countries among different constructs of interests (Lietz et al., 2017). To guarantee meaningful comparison, two features of the study design are of importance. The first is concerned with whether the same instrument was used across all participants to collect responses. In ILSA studies, by design, the battery of items are equivalent, besides their translation variants (Schulz & Carstens, 2020). The second relates to the sampling design. For example, the International Civic and Citizenship Education Study uses a stratified probability sample of schools, and select intact classrooms within selected schools, to represent the population of 8th graders in the participating countries (Schulz et al., 2018). In this design, the observations are weighted, and the secondary user can include the complex sampling design information to generate estimates that are generalizable to the target population.

Monte Carlo simulations of Zheng and Yang (2016) showed that including the complex sampling design information in the estimation of the response model produced less biased estimates. Thus, including the complex sampling design allows us to produce results generalizable to the population of students using latent variable models (Stapleton, 2013; Sterba, 2009). In contrast, a response model fitted to the same data, while ignoring its complex sampling design, would be assuming observations that come from a simple random sample. This means that the estimated parameters would not properly refer to the target population from which the observations were collected. The inclusion of the survey complex sampling design within the response model estimation allows us to correctly interpret and generalize the estimates of the response model to the intended target population.

In this paper, we use the “Students’ experience of physical and verbal abuse at school” from ICCS 2016 to illustrate how to fit a continuation ratio response model and interpret its estimates referring to the target population of the study. In the following section, we describe the data source, the measures, and the expansion technique used to specify the continuation ratio response model.

4 Methods

4.1 Selected Data and Measures for Illustrations

We used the responses from the International Civic and Citizenship Education Study from ICCS 2016. We retrieved the responses from the Latin American countries that participated in the study, including Chile, Colombia, Dominican Republic, Mexico, and Peru. All these countries are Spanish speaking countries; thus, the responses we selected for illustration purposes were generated with the same battery of items. ICCS 2016 uses a two-stage sampling design, where schools are chosen using a stratified design in each participating country, and in a second step, all students from the same classroom are selected to participate in the study. With this sampling design, ICCS 2016 reaches representative samples of 8th graders for all participating countries. Nominal samples are composed of 5081 students and 178 schools/classrooms from Chile, 5609 students and 150 schools/classrooms from Colombia, 3937 students and 141 schools/classrooms from Dominican Republic, 5526 students and 213 schools/classrooms from Mexico, and 5166 students and 206 schools/classrooms from Peru.

Dependent variables The “Students’ experience of physical and verbal abuse at school” (y_{pi}) scale. Students indicate how frequent six bullying events have happened to them in the last 3 months, using four ordered category options: “none at all”, “once”, “two to four times”, and “five times or more”. Examples of these bullying events are “A student said things about you to make others laugh” and “A student threatened to hurt you”. Table 2 shows the full list of items and the content of the scale, including the frame of reference and its response options.

4.2 Data Expansion Technique

As we argued earlier, not all the off-the-shelf software can fit a continuation ratio model in decreasing order and account for the complex sampling design in its estimates. Therefore, to tackle the first issue and make fitting the CRM a possibility, we follow the expansion technique proposed by Gochyyev (2015). Under this approach, the original item with k responses is expanded into $k-1$ pseudo items into a wide data format, where the observed response marks the suffered event in

Table 2 “Students’ experience of physical and verbal abuse at school” in ICCS 2016

Frame	During the last 3 months, how often did you experience the following situations at your school? (Please tick only one box in each row.)				
	Not at all	Once	2–4 times	5 times or more	
bul1	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
bul2	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
bul3	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
bul4	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
bul5	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
bul6	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

Note: “Students’ experience of physical and verbal abuse at school” scale contents, including the frame of reference, each item and its response options (see Köhler et al., 2018, p. 10). In the present manuscript we use “bul1-bul6” to refer to each specific item

Table 3 Illustration of the conversion of the responses to the polytomous item *bul1* from the bullying scale, into three pseudo items

Response	bul1	bul1_3	bul1_2	bul1_1	n
Not at all	0	0	0	0	9994
Once	1	0	0	1	7495
2–4 times	2	0	1	NA	3521
5 times or more	3	1	NA	NA	3804
Missing response	NA	NA	NA	NA	505

Note: Response = is the response category. *bul1* = refers to the original responses to item “A student called you by an offensive nickname.”, *bul1_3*- *bul1_1* = are the dummy coded values following the decreasing ordered expansion technique proposed by Gochyyev (2015). *n* = is the observed number of cases in each category

the respective frequency category with a dummy coded variable. For example, if a student answers that he/she has been called by an offensive nickname in the last 3 months once (*bul1*), then we need a pseudo item (*bul_1*) coded as one, while the rest of response options is left as missing. In contrast, if the student answers “2 to 4 times”, this response option is coded as 1, the previous categories are coded as 0, and the higher categories are left as missing. Finally, if the student answered, “5 times or more”, then “*bul_3*” is coded as 1 for this category and all previous categories are dummy coded as zero. Table 3 shows a schematic representation of this procedure.

We proceed in the same way with all the original items, generating a dummy coded matrix of values with three additional vectors for each original item, thus creating a matrix of *n* cases, and 18 new columns.

The expansion technique implemented here is similar to transforming a person response data, to a multivariate format by including the generated pseudo items as additional columns (Hoffman, 2015). Person response data uses a single row per person, and a single column for each ordinal response. Transforming the person response data to the multivariate format is similar to the approach used by Masyn (2009, 2014), in survival models, where single records are turn into multivariate data frames by including the needed dummy coded variables as additional columns in a data frame, to generate continuation ratio logits for a single survival spell. In contrast, De Boeck and colleagues (2011; Jeon & De Boeck, 2016) relied on a stack data format expansion to generate estimates for continuation ratio logits. In the latter case, the person record and its responses are expanded as rows instead of columns and includes additional rows per response as needed. The multivariate format allows retrieving the same estimates that the stack data format with appropriate model constraints (e.g., slopes constraints to one, and free variance); thus, conforming to a Rasch Model specification over the generated pseudo items. We rely on the multivariate format so we can use the complex sampling weights and stratification variables into the estimation without further transformations of these latter variables. Moreover, the generated pseudo items are allowing us to retrieve the expected continuation ratio logits of interest.

4.3 Analytical Strategy

To consider the survey complex sampling design, we prepared the original data to merge all observations into a single file, thus generating a pooled international sample. We first generate normalized weights, also called senate weights (Gonzalez, 2012). These are the total survey weights, linearly transformed to weight cases up to a constant, in this case up to 1000. This procedure assures that cases from different countries weight observations to a common figure. If this procedure is not undertaken, then estimates can be distorted since survey weights expand observations up to the expected population. The expected population varies between countries. A country like Mexico for example, expands its observed sample of 5526 cases to an expected population of 2,240,779 students in 8th grade; while the sample from Dominican Republic, expands its sample from 3937 cases to an expected count of 138,418 of 8th grade students. As such, the original sampling weights would give a higher contribution to Mexico, than to Dominican Republic observations. Thus, if a secondary user wants to generate a pooled sample to fit models, where two or more samples are contained, survey weights need to be re-scaled accordingly. We called the scaled survey weight variable “ws”. Similarly, we converted the pseudo strata vector from each country that varies from 1 to 75, into a unique vector across countries thus generating 375 unique pseudo strata (“id_s”). We proceed in a similar way with the cluster variable, so each value identifies each primary sampling unit, i.e., the schools, as a unique value across countries (“id_j”).

We use Taylor Series Linearization to get corrected standard errors and pseudo maximum likelihood (Asparouhov, 2005) as implemented in Mplus 8.6 (Muthén & Muthén, 2017). Table 4 summarizes the code used to fit the chosen model.

5 Results

In this section we describe the results for the fitted model, separated in two parts. We first describe the fixed effect estimates of the model, used to model item properties. Next, we describe the random effects estimates used to model person estimates in the model. To summarize the main results, we use a item-person map as shown in Fig. 1.

5.1 Item Side

A response model resembles a multivariate analysis of variance. In this light, the item parameters can tell us how frequent an event of bullying is across the population of students, accounting for the rest of bullying events present in the battery, and the propensity of students being bullied at school. For the present application,

Table 4 Mplus syntax used to fit the continuation ratio response model with decreasing ordered

<pre> TITLE:CRM_MLR; DATA: FILE = "bull_scale.dat"; VARIABLE: NAMES = id_i id_j id_s ws bul1 bul2 bul3 bul4 bul5 bul6 bul1_2 bul1_1 bul1_0 bul2_2 bul2_1 bul2_0 bul3_2 bul3_1 bul3_0 bul4_2 bul4_1 bul4_0 bul5_2 bul5_1 bul5_0 bul6_2 bul6_1 bul6_0; MISSING=.; CATEGORICAL = bul1_2 bul1_1 bul1_0 bul2_2 bul2_1 bul2_0 bul3_2 bul3_1 bul3_0 bul4_2 bul4_1 bul4_0 bul5_2 bul5_1 bul5_0 bul6_2 bul6_1 bul6_0 ; USEVARIABLES = bul1_2 bul1_1 bul1_0 bul2_2 bul2_1 bul2_0 bul3_2 bul3_1 bul3_0 bul4_2 bul4_1 bul4_0 bul5_2 bul5_1 bul5_0 bul6_2 bul6_1 bul6_0 ; IDVARIABLE = id_i; </pre>	<pre> WEIGHT = ws; CLUSTER = id_j; STRATIFICATION = id_s; ANALYSIS: TYPE = COMPLEX; ESTIMATOR = MLR; MODEL: [theta@0]; theta; theta by bul1_0@1; theta by bul1_1@1; theta by bul1_2@1; theta by bul2_0@1; theta by bul2_1@1; theta by bul2_2@1; theta by bul3_0@1; theta by bul3_1@1; theta by bul3_2@1; theta by bul4_0@1; theta by bul4_1@1; theta by bul4_2@1; theta by bul5_0@1; theta by bul5_1@1; theta by bul5_2@1; theta by bul6_0@1; theta by bul6_1@1; theta by bul6_2@1; OUTPUT: STANDARDIZED CINTERVAL RESIDUAL; SAVEDATA: SAVE = FSCORES; FILE = crm_svy_eap.dat; </pre>
--	---

Note: The Mplus syntax consists of a single column of text. In the present table this is presented in two columns to make the code fit into a single page, starting from left to right

the population is the pooled population of students from Chile, Colombia, Mexico, Dominican Republic and Peru.

In these results we can see that the most frequent bullying event across students is to be mocked (bul2) and having an offensive nickname (bul1). In the present model a value of 0 logits can be interpreted as the average propensity to be bullied by students. With that in mind we can see in the figure that being mocked and having an offen-sive nickname are two events that can be commonly experienced by students who are exposed to an average chance or above chance to be bullied, which in this case represent more than half of the students. The least frequent event is being shamed on the internet by other students (bul6). According to the models' results, only students in the highest propensities to be bullied are likely to suffer

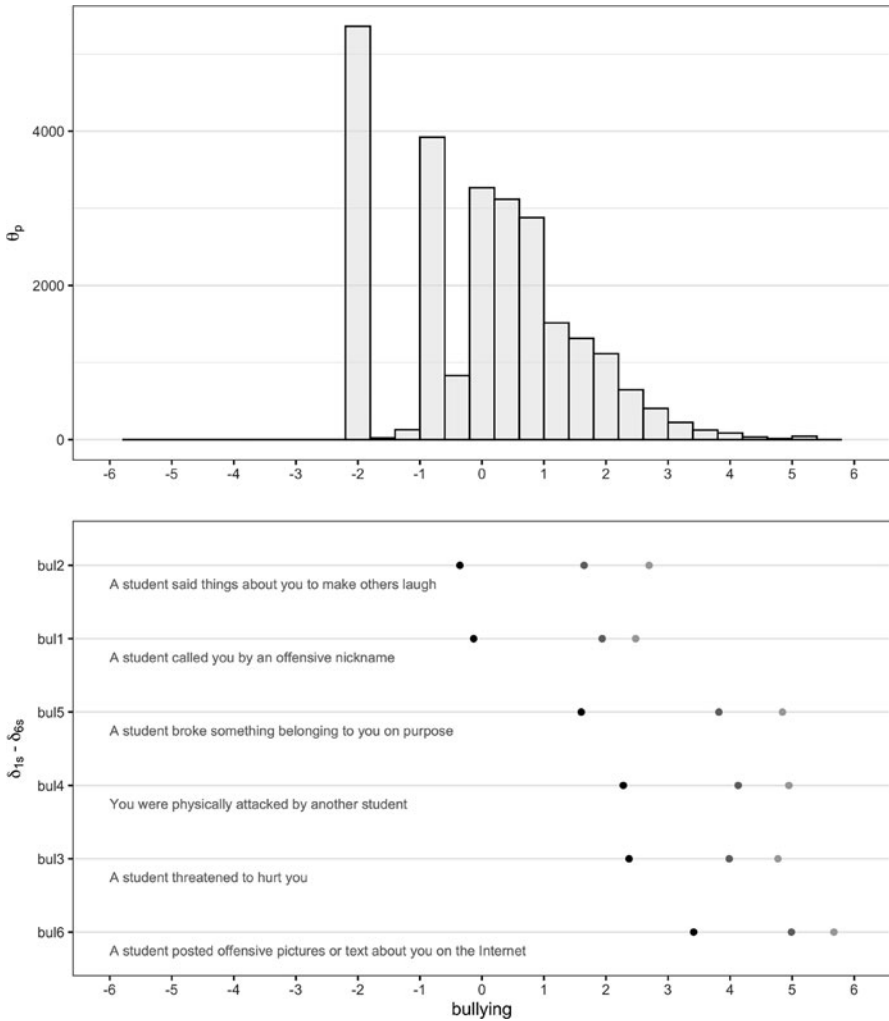


Fig. 1 Item person map of the bullying scale, using the estimates of the continuation response model

this kind of abuse. The estimated logit of this last item, for its first odds comparing “once” over “not at all” responses, is higher than three logits from the center of the distribution. Students with such a high propensity, and therefore likely to have suffered from being shamed upon on the internet by their peers in the last 3 months are also at higher risk of suffering all the other kinds of bullying, including threats and physical attacks.

5.2 *Person Side*

The generated scores (EAP predictions) presents a person separability of .72 (Adams, 2005). This observed reliability index is slightly smaller in contrast to other background and context questionnaires (Schulz et al., 2018). However, this result is not uncommon given that the distribution of cases is positively skewed. This distribution is common for measures of bullying, where there is a high proportion of students who do not report suffering from bullying. Thus, mimicking a case of a “zero-inflation”, where a substantive portion of the observations do not present the attribute of interest (i.e., suffering from bullying), while the rest of the distribution becomes smaller at higher values of the propensity continuum.

We fit a CRM, GRM, and PCM models to the same data, estimating person and items parameters only (θ , δ), leaving item slopes fixed to a unity, and including the complex sampling design (as seen in Table 4). Although, the relative fit indices favor the PCM model ($BIC_{pcm} = 226970.254$; $BIC_{grm} = 227050.074$; $BIC_{crm} = 227746.389$), the generated score per person (EAP predictions), are highly correlated between models: PCM with GRM ($r = .99$); PCM with CRM ($r = .99$); GRM with CRM ($r = 1.00$). We also correlated the generated scores of the CRM model, with a sum score, which also presents a high correlation ($r = .95$). We also estimated an informative index (in Asparouhov, 2006) for including the sampling weights EAP estimates means and theta variance. This informative index compares an unweighted model parameter, and its weighted model parameter counterpart divided by its variance or standard errors (using a t-test). These two comparisons yield t-scores close to zero. Thus, the sampling survey design of this study, is not informative for the fitted model. So, in general terms, we can conclude that students can be ordered in terms of their propensity to be bullied using the CRM model in a manner comparable with other, more commonly used, polytomous models.

6 Conclusion and Discussion

We have illustrated a method to fit a continuation ratio model subject to a complex sampling design, commonly used in ILSA. Using the expansion technique proposed by Gochyyve (2015) we have shown how it is possible to fit the continuation ratio response models using statistical software that can fit Rasch models and that can deal with missing data. Using this technique, it is possible to produce complex sampling design estimates, generalizable to the finite population. Kim (2016) has illustrated the applicability of the continuation ratio model to polytomous ordered responses. Calderón and González (2021) compared the performance of some traditional polytomous IRT models with the more recently introduced IRTree models in the modeling of self-report non-cognitive latent traits. Such comparison is motivated by the fact that IRTree models allow to account for extreme response style (ERS)

effects in attitudinal measurements, while at the same time providing estimates of the target trait. The present work aims to expand these applications to the context of large-scale assessment studies, where complex sampling design is an informative feature to produce scores and results.

The continuation ratio response model, can be expanded to include floor effects (Yamamoto, 1987), or “censored” like latent classes (Masyn, 2003). This is important for the school victimization literature. Bullying scales generated with sum scores, present a substantive proportion of cases at the lowest possible value, thus identifying cases who declare not suffering any form of bullying at school. Accordingly, previous studies have used zero inflated poisson regression to condition sum scores on predictors (e. g., Rutkowski et al., 2013). Censored cases are cases that are at the lowest level of bullying risk. In the survival literature, censored cases are worrisome. If this type of cases is not accounted for in the model, inferences regarding the relationship with covariates are downward biased (Masyn, 2003). This latter bias is problematic for the study of the effectiveness of interventions on school bullying, the comparison of schools regarding bullying prevalence, and the general study of risk and protective factors of bullying. Thus, impeding one of the mains aim of ILSA studies: to provide valid and reliable information for public policy.

Kim (2016) suggests that the continuation ratio response model present advantages for metric linking and equating. In essence, once pseudo items are generated users can easily use different routines available for dichotomous items, without large modifications. Moreover, invariance studies using DIF techniques (Masyn, 2017), different multigroup models (Van de Vijver et al., 2019) and alignment methods (Muthén & Asparouhov, 2014) can be easily adapted using the expansion technique illustrated in the present work.

The informative index proposed by Asparouhov (2006) shows that the sampling design of ICCS 2016 is not substantively informative for the studied responses of the bullying scale. As such, a clustered error model could be enough to fit the proposed model, and get corrected standard errors (Stapleton, 2013). However, even though the bullying scale generated estimates that are similar with or without sampling design, it is not advisable to plainly ignore the study survey design for any other context questionnaire scale. Complex sampling design in large scale assessment includes stratification variables of schools, that are informative for the target score of the test (Meinck, 2020), in the case of ICCS the civic knowledge test. Moreover, the stratification variables used to select schools could be informative for the battery of items of socioeconomic status. As such, multi-item scale scores related to the stratification variables of each country could display higher informative indexes. Yet, the informative indexes calculated for the bullying scale mean and variance were quite small.

Further research is needed to illustrate the expected benefits of the presented model for scale linking, DIF studies, and invariance evaluation. Specially, for the expected benefits of rethinking the response model for inferences with censored like cases.

Acknowledgements Research funded by the Fondo Nacional de Desarrollo Científico y Tecnológico FONDECYT N° 1201129 and FONDECYT N° 11180792. David Torres-Irribarra and Jorge González were partially supported by the Agencia Nacional de Investigación y Desarrollo (ANID) Millennium Nucleus on Intergenerational Mobility: From Modelling to Policy (MOVI); (NCS2021072).

References

- Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, 31(2–3), 162–172. <https://doi.org/10.1016/j.stueduc.2005.05.008>
- Agresti, A. (2019). *An introduction to categorical data analysis* (3rd ed.). John Wiley & Sons, Inc..
- Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 12(3), 411–434. https://doi.org/10.1207/s15328007sem1203_4
- Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics – Theory and Methods*, 35(3), 439–460. <https://doi.org/10.1080/03610920500476598>
- Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1). <https://doi.org/10.18637/jss.v067.i01>
- Calderón, F., & González, J. (2021). Polytomous IRT models versus IRTree models for scoring non-cognitive latent traits. In M. Wiberg, D. Molenaar, J. González, U. Böckenholt, & J. Kim (Eds.), *Quantitative psychology* (pp. 113–125). https://doi.org/10.1007/978-3-030-74772-5_11
- Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6). <https://doi.org/10.18637/jss.v048.i06>
- De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, 48(Code Snippet 1). <https://doi.org/10.18637/jss.v048.c01>
- De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, 39(12), 1–28. <https://doi.org/10.18637/jss.v039.i12>
- Gochyyev, P. (2015). *Essays in psychometrics and behavioral statistics*. University of California. <https://escholarship.org/content/qt36b8p5nw/qt36b8p5nw.pdf>
- Gonzalez, E. J. (2012). Rescaling sampling weights and selecting mini-samples from large-scale assessment databases. *IERI Monograph Series Issues and Methodologies in Large-Scale Assessments*, 5, 115–134.
- Hemker, B. T., Andries van der Ark, L., & Sijtsma, K. (2001). On measurement properties of continuation ratio models. *Psychometrika*, 66(4), 487–506. <https://doi.org/10.1007/BF02296191>
- Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. Psychology Press. <http://www.piles-of-variance.com/>
- Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods*, 48(3). <https://doi.org/10.3758/s13428-015-0631-y>
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2), 183–202. <https://doi.org/10.1007/BF02289343>
- Kim, S. (2016). Continuation ratio model in item response theory and selection of models for polytomous items. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang (Eds.), *Quantitative psychology research* (Vol. 196, pp. 1–13). Springer International Publishing. https://doi.org/10.1007/978-3-319-38759-8_1
- Köhler, H., Weber, S., Brese, F., Schulz, W., & Carstens, R. (2018). In H. Köhler, S. Weber, F. Brese, W. Schulz, & R. Carstens (Eds.), *ICCS 2016 user guide for the international database*. International Association for the Evaluation of Educational Achievement (IEA).
- Lietz, P., Cresswell, J. C., Rust, K., & Adams, R. J. (2017). In P. Lietz, J. C. Cresswell, K. Rust, & R. J. Adams (Eds.), *Implementation of large-scale education assessments*. John Wiley & Sons, Ltd.. <https://doi.org/10.1002/9781118762462>

- Martin, M.O., Mullis, I.V.S., Hooper, M.: Methods and procedures in PIRLS 2016. (2017). <https://timssandpirls.bc.edu/publications/pirls/2016-methods.html>
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174. <https://doi.org/10.1007/BF02296272>
- Masyn, K. E. (2003). *Discrete-time survival mixture analysis for single and recurrent events using latent variables*. University of California Los Angeles. <http://www.statmodel.com/download/masyndissertation.pdf>
- Masyn, K. E. (2009). Discrete-time survival factor mixture analysis for low-frequency recurrent event histories. *Research in Human Development*, 6(2–3), 165–194. <https://doi.org/10.1080/15427600902911270>
- Masyn, K. E. (2014). Discrete-time survival analysis in prevention science. In Z. Sloboda & H. Petras (Eds.), *Defining prevention science* (pp. 513–535). Springer US. <https://doi.org/10.1007/978-1-4899-7424-2>
- Masyn, K. E. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(2), 180–197. <https://doi.org/10.1080/10705511.2016.1254049>
- Meinck, S. (2020). Sampling, weighting, and variance estimation. In H. Wagemaker (Ed.), *Reliability and validity of international large-scale assessment* (pp. 113–129). Springer International Publishing. https://doi.org/10.1007/978-3-030-53081-5_7
- Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology*, 5(AUG), 1–7. <https://doi.org/10.3389/fpsyg.2014.00978>
- Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.
- OECD. (2019). *PISA 2018 results: What school life means for students' lives* (Vol. III). PISA, OECD Publishing. <https://doi.org/10.1787/acd78851-en>
- OECD. (2020). Scaling procedures and construct validation of context questionnaire data. In *PISA 2018 technical report* (pp. 1–39). OECD.
- Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). GLLAMM manual. In *The Berkeley Electronic Press (bepress)* (No. 160). The Berkeley Electronic Press (bepress). <http://www.bepress.com/ucbbiostat/paper160>
- Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Research*, 39(2), 142–151. <https://doi.org/10.3102/0013189X10363170>
- Rutkowski, D., Rutkowski, L., & Wild, J. (2013). Predictors of school violence internationally: The importance of immigrant status and other factors. In *5th IEA international research conference, 26–28 June 2013*. http://www.iea.nl/fileadmin/user_upload/IRC/IRC_2013/Papers/IRC-2013_Rutkowski_etal.pdf
- Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *ETS Research Bulletin Series*, 1968(1), i–169. <https://doi.org/10.1002/j.2333-8504.1968.tb00153.x>
- Schulz, W., & Carstens, R. (2020). Questionnaire development in international large-scale assessment studies. In H. Wagemaker (Ed.), *Reliability and validity of international large-scale assessment* (Vol. 10, pp. 61–83). Springer International Publishing. https://doi.org/10.1007/978-3-030-53081-5_5
- Schulz, W., Carstens, R., Losito, B., & Fraillon, J. (2018). In W. Schulz, R. Carstens, B. Losito, & J. Fraillon (Eds.), *ICCS 2016 technical report*. International Association for the Evaluation of Educational Achievement (IEA).
- Stapleton, L. M. (2013). Incorporating sampling weights into single- and multilevel analyses. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large scale assessment: Background, technical issues, and methods of data analysis* (pp. 363–388). Chapman and Hall/CRC.
- Sterba, S. K. (2009). Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration. *Multivariate Behavioral Research*, 44(6), 711–740. <https://doi.org/10.1080/00273170903333574>

- Tutz, G. (1990). Sequential item response models with an ordered response. *The British Journal of Mathematical and Statistical Psychology*, 43(1), 39–55. <https://doi.org/10.1111/j.2044-8317.1990.tb00925.x>
- Tutz, G. (2016). Sequential models for ordered responses. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 1, pp. 139–151). CRC Press. <https://doi.org/10.1201/9781315374512>
- Van de Vijver, F. J. R., Avvisati, F., Davidov, E., Eid, M., Fox, J.-P., Le Donne, N., Lek, K., Meuleman, B., Paccagnella, M., & van de Schoot, R. (2019). Invariance analyses in large-scale studies. *OECD Education Working Papers*, 201, 1–110. <https://doi.org/10.1787/19939019>
- Van Der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement*, 25(3), 273–282. <https://doi.org/10.1177/01466210122032073>
- Vermunt, J. K., & Magidson, J. (2016). *Technical guide for latent GOLD choice 5.1: Basic, advanced and syntax*. Statistical Innovations Inc..
- Yamamoto, K. (1987). *A model that combines IRT and latent class models*. University of Illinois at Urbana-Champaign. <http://ezproxy.puc.cl/dissertations-theses/model-that-combines-irt-latent-class-models/docview/303564885/se-2?accountid=16788>
- Yin, L., & Fishbein, B. (2020). Creating and interpreting the TIMSS 2019 context questionnaire scales. In M. O. Martin, M. Von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report*. TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA).
- Zheng, X., & Yang, J. S. (2016, January). Using sample weights in item response data analysis under complex sample designs. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative psychology research* (pp. 123–137). Springer International Publishing. https://doi.org/10.1007/978-3-319-38759-8_10

Impact of Likelihoods on Class Enumeration in Bayesian Growth Mixture Modeling



Xin Tong , Seohyun Kim , and Zijun Ke 

Abstract Bayesian methods have been widely used to estimate models with complex structures. To assess model fit and compare different models, researchers typically use model selection criteria such as Deviance Information Criteria (DIC), Watanabe-Akaike Information Criteria (WAIC) and leave-one-out cross validation (LOO-CV), the calculation of which is based on the likelihoods of the models. When models contain latent variables, the likelihood is often specified as conditional on the latent variables in popular Bayesian software (e.g., BUGS, JAGS, and Stan). Although this practice reduces computation work and does not affect model estimation, the previous literature has shown that model comparisons based on the conditional likelihood could be misleading. In contrast, marginal likelihoods can be obtained by integrating out the latent variables and be used to calculate model selection criteria. In this study, we evaluate the effect of using conditional likelihoods and marginal likelihoods in model selection for growth mixture models. Simulation results suggest that marginal likelihoods are much more reliable and should be generally used for growth mixture modeling.

Keywords Marginal likelihood · Conditional likelihood · DIC · WAIC · LOO-CV

X. Tong (✉)

Department of Psychology, University of Virginia, Charlottesville, VA, USA
e-mail: xt8b@virginia.edu

S. Kim

Kaiser Permanente Mid-Atlantic Permanente Research Institute, Rockville, MD, USA
e-mail: sonia.s.kim@kp.org

Z. Ke

Department of Psychology, Sun Yat-sen University, Guangzhou, China
e-mail: keziyun@mail.sysu.edu.cn

1 Introduction

Growth mixture modeling (GMM) is a method for identifying multiple unobserved subgroups in a population, describing longitudinal change within each subgroup, and examining differences in change among those subgroups (Ram & Grimm, 2009). GMM has been increasingly used in social and behavioral sciences (e.g., Frankfurt et al., 2016; McDermott et al., 2018; Smith & Ehlers, 2020; Ren et al., 2021) to flexibly model growth trajectories with substantial individual variations. Despite the popularity of GMM, several issues are involved in its model estimation (Bauer & Curran, 2003; Hipp & Bauer, 2006), including violation of distributional assumptions, obtaining local solutions, nonconvergence, etc. Researchers have made efforts in addressing these issues, many of which are in the Bayesian framework (e.g., Depaoli, 2013; Kim et al., 2021a; Lu et al., 2011). Bayesian approaches are relatively flexible in accounting for the nonnormality in data and enable incorporating prior information into model estimation to help yield converged results when there are not enough samples or latent classes are not well separated (Depaoli, 2014; Kim et al., 2021b). In addition, data augmentation and Markov chain Monte Carlo (MCMC) techniques can be naturally applied in the Bayesian framework to reduce the mathematical demands for complex model estimation.

Deciding the appropriate number of latent classes (i.e., unobserved subgroups) is critical in GMM and is typically achieved by comparing models with different number of latent classes and selecting the best fitting model. In Bayesian statistics, model comparison and selection can be performed using the Bayes factor which is the ratio of the posterior odds to the prior odds of two competing models. Since the calculation of the Bayes factor is often difficult and greatly influenced by the priors, model comparison in GMM is typically conducted using information criteria and cross validation which estimate out-of-sample predictive accuracy using within-sample fits. The calculation of the model selection criteria is based on the likelihoods of the models.

In popular Bayesian software (e.g., BUGS, JAGS, and Stan), the likelihood of GMM is often specified as conditional on the latent variables. However, recent studies (e.g., Kim et al., 2021a; Merkle et al., 2019) reported that model selection and comparison based on the conditional likelihood in latent variable modeling can be misleading. Instead, marginal likelihoods were used where latent variables were integrated out in the likelihood functions. Although conditional and marginal likelihoods do not make differences in terms of model estimation after all Markov chains converge, the distinction between them in model selection is substantial but is often overlooked. As far as we are aware, only Merkle et al. (2019) has particularly studied the difference between conditional and marginal likelihoods and recommended use of marginal likelihood based information criteria in Bayesian latent variable analysis.

Due to the complexity of GMM and unique challenges associated with it, in this study, we will evaluate the performance of conditional and marginal likelihood in GMM class enumeration. We focus on two information criteria: Deviance Information Criterion (DIC; Spiegelhalter et al., 2002) and Widely Applicable Information Criterion (Watanabe-Akaike Information Criterion, WAIC; Watanabe, 2010), and one cross validation approach: leave-one-out cross validation (LOO-CV; Gelman et al., 2013; Vehtari et al., 2017). Their performance based on different likelihoods in GMM class enumeration will be investigated. The paper is organized as follows. We first briefly review growth mixture models, introduce the associated conditional and marginal likelihoods, and different model selection criteria. Then we use a simulation study to assess the impact of conditional and marginal likelihoods on GMM model selection. Recommendations are provided at the end of the article.

2 Bayesian GMM Model Selection

2.1 A Brief Review of Growth Mixture Models

Growth mixture models extend growth curve models by assuming that a population consists of a number of latent classes (i.e., unobserved subgroups) and each latent class is characterized by a unique growth trajectory. Suppose that a population consisted of G latent classes that have distinct patterns of change. Let $\mathbf{y}_i = (y_{i1}, \dots, y_{iT_i})'$ denote a vector of T_i repeated observations for individual i ($i \in \{1, \dots, N\}$). A general form of growth mixture models can be expressed as

$$\mathbf{y}_i | (z_i = g) = \mathbf{A}_i \mathbf{b}_{ig} + \boldsymbol{\epsilon}_i, \quad (1)$$

where the subscript g indicates that a corresponding parameter or variable is class-specific. In this model, z_i represents a class indicator for individual i with mixing proportion for class g being $P(z_i = g) = \pi_g$, \mathbf{A}_i is a $T_i \times q$ matrix of factor loadings that determines the shape of the growth trajectories, \mathbf{b}_{ig} is a $q \times 1$ vector of latent factors for class g ($g \in \{1, \dots, G\}$), and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{iT_i})'$ is a $T_i \times 1$ vector of intraindividual measurement errors. The latent factors are often assumed to follow multivariate normal distributions such that $\mathbf{b}_{ig} \sim MN(\boldsymbol{\beta}_g, \boldsymbol{\Psi}_g)$, where $\boldsymbol{\beta}_g$ is the mean of \mathbf{b}_{ig} and $\boldsymbol{\Psi}_g$ is the covariance matrix of \mathbf{b}_{ig} . The measurement errors are also assumed to be normally distributed, $\boldsymbol{\epsilon}_i \sim MN(\mathbf{0}, \boldsymbol{\Sigma}_g)$, leading the conditional mean of \mathbf{y}_i given \mathbf{b}_{ig} to be $E(\mathbf{y}_i | \mathbf{b}_{ig}) = \mathbf{A}_i \mathbf{b}_{ig}$. In practice, it is common to further assume that the intraindividual measurement errors have equal variances and are independent across time, so that $\boldsymbol{\Sigma}_g = \sigma_g^2 \mathbf{I}$, where σ_g^2 is a scale parameter for class g . We assumed this measurement error structure for the rest of this study.

2.2 Conditional and Marginal Likelihoods of Growth Mixture Models

With the normality assumption, the likelihood function of the model in Eq. (1) can be specified. As stated previous, popular Bayesian software often specify the likelihood as conditional on the latent variables. That is,

$$L_C(\mathbf{b}_{ig}, z_i, \sigma_g^2 | \mathbf{y}) = p(\mathbf{y} | \mathbf{b}_{ig}, z_i = g, \sigma_g^2), \quad (2)$$

where $p(\mathbf{y}_i | \mathbf{b}_{ig}, z_i = g, \sigma_g^2)$ is the density function of the multivariate normal distribution $MN(\mathbf{A}_i \mathbf{b}_{ig}, \sigma_g^2 \mathbf{I})$.

To obtain the marginal likelihood of Model (1), the latent variables \mathbf{b}_{ig} and z_i have to be integrated out of the conditional likelihood. The marginal likelihood for the normal-distribution-based GMM has a closed form:

$$L_M(\boldsymbol{\beta}_g, \boldsymbol{\Psi}_g, \pi_g, \sigma_g^2 | \mathbf{y}) = p(\mathbf{y} | \boldsymbol{\beta}_g, \boldsymbol{\Psi}_g, \pi_g, \sigma_g^2) = \prod_{i=1}^N \sum_{g=1}^G \pi_g p(\mathbf{y}_i | \boldsymbol{\beta}_g, \boldsymbol{\Psi}_g, \sigma_g^2), \quad (3)$$

where $p(\mathbf{y}_i | \boldsymbol{\beta}_g, \boldsymbol{\Psi}_g, \sigma_g^2)$ is the density function of the multivariate normal distribution $MN(\mathbf{A}_i \boldsymbol{\beta}_g, \mathbf{A}_i \boldsymbol{\Psi}_g \mathbf{A}_i' + \sigma_g^2 \mathbf{I})$.

Given the likelihood functions, model selection criteria can be computed. Based on the model selection criteria, we can compare GMMs with different number of latent classes and select the best fitting model.

2.3 Model Comparison Criteria

In this paper, we use DIC, WAIC, and LOO-CV to select the optimal number of latent classes for GMM. We briefly introduce the three model comparison criteria below.

DIC was proposed by Spiegelhalter et al. (2002). Although it has received much criticism (e.g., Celeux et al., 2006), it is widely used in Bayesian model selection. DIC is defined as the sum of the expected deviance over the parameter space and the effective number of model parameters,

$$DIC = \bar{D} + p_D.$$

The expected deviance is

$$\bar{D} = E_{\boldsymbol{\Theta}}[-2 \log p(\mathbf{y} | \boldsymbol{\Theta}) | \mathbf{y}] + C,$$

where $\boldsymbol{\Theta}$ is a set of model parameters, and C is a constant that can be canceled out when comparing models. \bar{D} is calculated as the posterior mean of the deviance. The

effective number of parameters, p_D , measures the complexity of the model and is defined as

$$p_D = \bar{D} - \hat{D},$$

where \hat{D} is the deviance calculated at the posterior mean of Θ . Models with smaller DICs are preferred.

WAIC was proposed more recently and have been shown to have advantages over DIC (Vehtari et al., 2017). WAIC uses the entire posterior distribution, is asymptotically equal to Bayesian cross validation, is invariant to parameterization, and works for singular models. We used the following definition of WAIC (Gelman et al., 2013).

$$WAIC = -2 \sum_{i=1}^N \log \left(\frac{1}{S} \sum_{s=1}^S p(y_i | \Theta^{(s)}) \right) + 2 \sum_{i=1}^N Var_{s=1}^S \log p(y_i | \Theta^{(s)}),$$

where S is the number of MCMC iterations, $\Theta^{(s)}$ is a draw from the posterior distribution at the s th iteration, and $Var_{s=1}^S$ represents the sample variance, $Var_{s=1}^S a_s = \frac{1}{S-1} \sum_{s=1}^S (a_s - \bar{a})^2$. Models with smaller WAICs are preferred.

LOO-CV evaluates the model fit based on an estimate of the log predictive density of the hold-out data. Each data point is taken out at a time to cross validate the model that is fitted based on the remaining data. LOO-CV is defined as

$$LOO = -2 \sum_{i=1}^N \log \int p(y_i | \Theta) p(\Theta | y_{-i}) d\Theta,$$

and in practice, it can be approximately calculated as

$$\widehat{LOO} = -2 \sum_{i=1}^N \log \frac{1}{\frac{1}{S} \sum_{s=1}^S \frac{1}{p(y_i | \theta^{(s)})}}.$$

Vehtari et al. (2017) showed that although WAIC is asymptotically equal to LOO-CV, LOO-CV is more robust in the finite case with weak priors or influential observations.

3 A Simulation Study

We now present a simulation study to evaluate the impact of conditional and marginal likelihoods based model selection criteria on GMM class enumeration.

We generated data from a two-class linear growth mixture model with 4 equally spaced measurement occasions. Namely, in Eq. (1), $G = 2$, $T_i = 4$, $\mathbf{A}_i = ((1, 1, 1, 1)', (0, 1, 2, 3)')$, the latent intercept and slope for class 1, $\mathbf{b}_{i1} \sim MN(\boldsymbol{\beta}_1, \boldsymbol{\Psi}_1)$ and the latent intercept and slope for class 2, $\mathbf{b}_{i2} \sim MN(\boldsymbol{\beta}_2, \boldsymbol{\Psi}_2)$. The covariance matrix of the latent intercepts and slopes were set to be $\boldsymbol{\Psi}_g = \boldsymbol{\Psi} = \begin{pmatrix} 0.25 & 0 \\ 0 & 0.04 \end{pmatrix}$ for $g = 1$ and 2 , and the intraindividual measurement error variance was set at $\sigma^2 = 0.2$. These variance and covariance parameters were assumed to be the same across the two latent classes. We manipulated three factors that could potentially influence the performance of GMM in the simulation study: sample size, class separation, and class proportions. Two different sample sizes were considered ($N = 300$ or 500). Class separation was characterized using Mahalanobis distance, which can be calculated as $MD = \sqrt{(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)' \boldsymbol{\Psi}^{-1} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)}$, where $\boldsymbol{\beta}_1$ represented the means of latent intercepts and latent slopes for the first latent class, and $\boldsymbol{\beta}_2$ represented the means of latent intercepts and latent slopes for the second latent class. We evaluated the influence of a high class separation and a relatively low class separation. For the high separation, the first class had an average latent intercept of 2 and an average slope of 0.5, $\boldsymbol{\beta}_1 = (2, 0.5)'$, so in general, the scores were increasing over time. The second class had an average latent intercept of 1 and an average slope of 0, $\boldsymbol{\beta}_2 = (1, 0)'$, indicating that the overall trajectory was a flat line. This setting yielded a Mahalanobis distance $MD = 3.2$. For the low class separation, $\boldsymbol{\beta}_1 = (1.5, 0.5)'$ and $\boldsymbol{\beta}_2 = (1, 0)'$, which had $MD = 2.7$. The class proportions were set to be either unbalanced (25% from the first class and 75% from the second class) or balanced (50% from both latent classes).

For each simulation condition, 200 datasets were generated. For each dataset, we fit growth mixture models with one class ($G = 1$), two classes ($G = 2$), and three classes ($G = 3$). Bayesian estimation of GMM was conducted using JAGS with the *rjags* R package (Plummer, 2017). JAGS is a Bayesian data analysis program that uses MCMC algorithms (e.g., Gibbs sampler) for generating samples from the posterior distribution of model parameters. In JAGS, we obtained posterior samples of the model parameters by augmenting the latent variables (\mathbf{b}_{i_g} and z_i). With the sampled parameters and the likelihood of the model, DIC, WAIC, and LOO-CV can be calculated. Since the likelihood can be calculated in Equation (2) or Equation (3) when a model contains latent variables, DIC, WAIC, and LOO-CV were calculated based on the conditional likelihood and the marginal likelihood, separately. We then assessed the performance of the model comparison criteria based on different likelihoods in class enumeration.

The following priors were used for model inferences as these priors had little information about the parameters: $p(\boldsymbol{\beta}_g) = MN(0, 10^3 \times \mathbf{I})$ for $g = 1, \dots, G$, $p(\boldsymbol{\Psi}) = InvWishart(2, \mathbf{I}_2)$, $p(\sigma^2) = InvGamma(.01, .01)$, and $p(\boldsymbol{\pi}) \sim Dirichlet(10\mathbf{j}_G)$, where G is the total number of latent classes, and \mathbf{j}_G is a $G \times 1$ vector that has 1 for all components for $G > 1$. The number of MCMC iterations was set to 10,000, and the first half of the iterations were discarded for burn-in. Although our pilot study showed that the 10,000 iterations were enough for the

chains to converge, to guarantee convergence, we also allowed up to 10 different starting values for each model estimation to obtain converged results.

3.1 Results

Figure 1a–b summarize the model selection results based on DIC, WAIC, and LOO-CV when class proportions are 25% and 75%. For balanced classes, the relative

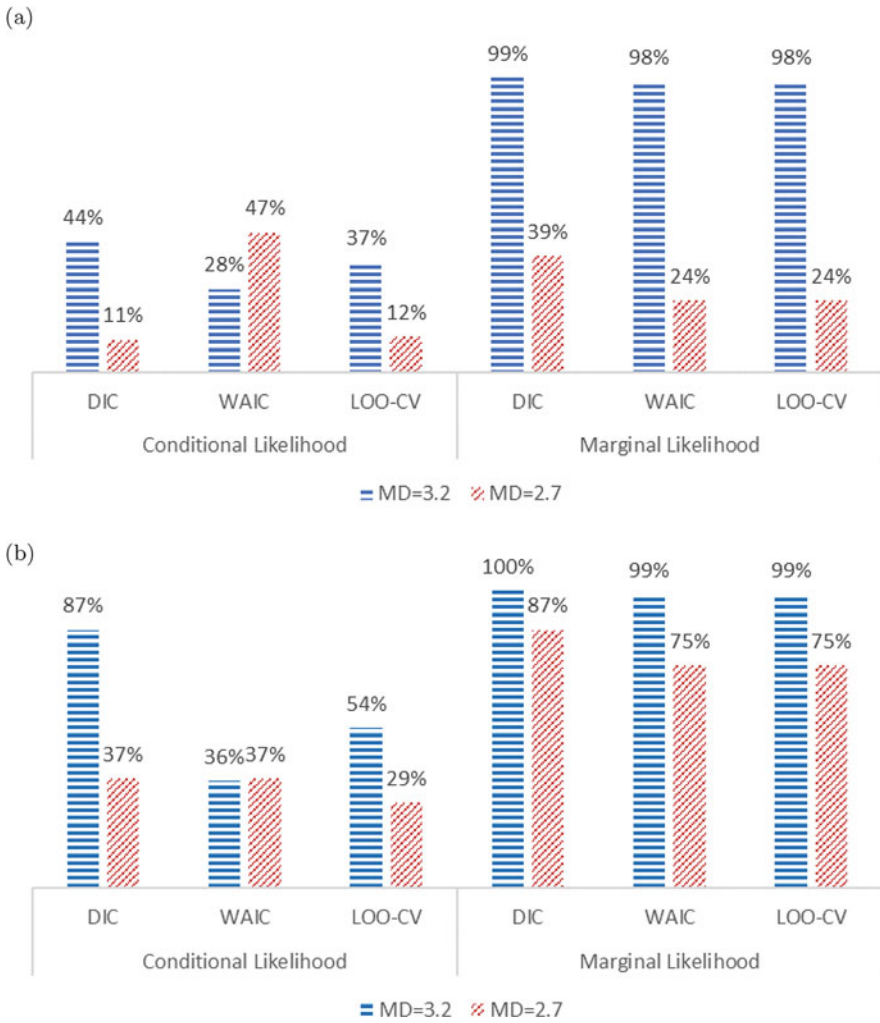


Fig. 1 Model selection results based on DIC, WAIC, and LOO-CV when class proportions are 25% and 75%. (a) $N = 300$. (b) $N = 500$

performance of the model comparison criteria based on conditional and marginal likelihoods has the same pattern and thus is not repeated in this section. Figure 1a and b report the results for $N = 300$ and $N = 500$, respectively. For each figure, the vertical axis (i.e., height of the bars) represents the probability of the correct model (2-class growth mixture model) being selected.

From Fig. 1, it can be seen that, in general, model selection based on DIC, WAIC, and LOO-CV is more likely to be correct when sample size is larger and the class separation is higher. When the class separation is relatively low (e.g., $MD = 2.7$), increasing sample size raises the chance to select the correct number of latent classes in GMM. For example, the marginal likelihood based DIC has 39% of the chances to select the correct model when $N = 300$. This percentage increases to 87% when the sample size is 500.

In addition, model selection criteria based on the marginal likelihood are more reliable as the bars on the right panel of Fig. 1a–b (marginal likelihood based criteria) are generally taller than the bars on the left panel of the figures (conditional likelihood based criteria). Although conditional likelihood based model selection criteria may perform well under some conditions (e.g., when $N = 500$, DIC calculated with the conditional likelihood has 87% of the chance to select the correct model), they are unstable in general. In practice, it is difficult to tell whether the conditional likelihood based model selection criteria are reliable or not for the study setting. Therefore, the conditional likelihood should not be used to calculate model selection criteria for model comparison. We have further investigated the probability of each model being selected for different data conditions. As demonstrated in Fig. 2, even when the class separation is relatively low, DIC, WAIC, and LOO-CV calculated based on the marginal likelihood almost always select the correct 2-class model. In contrast, the model selection criteria calculated based on the conditional likelihood tend to prefer simpler models (i.e., 1-class model) under this condition.

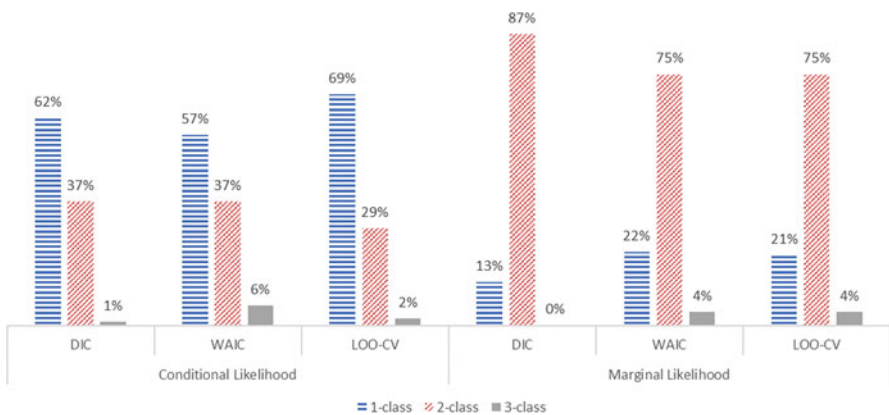


Fig. 2 The comparison between the conditional likelihood and marginal likelihood in selecting different growth mixture models when $N = 500$ and $MD = 2.7$. The height of the bar represents the percentage that the corresponding model is selected

Moreover, when the marginal likelihood is used, DIC, WAIC, and LOO-CV provide similar values. WAIC and LOO-CV, in particular, provide almost identical values. However, when the conditional likelihood is used, DIC, WAIC, and LOO-CV are not as similar as those we get from the marginal likelihood. In addition, WAIC and LOO-CV seem to perform slightly worse than DIC, especially when the class separation is relatively low.

4 Discussion

Bayesian methods have been increasingly used for GMM model estimation because of their flexibility and capability to handle model with complex structures. An important task of GMM is to determine the number of latent classes, and is typically conducted by model comparisons. Commonly used Bayesian model comparison criteria are calculated based on the likelihood of the model. In this paper, we evaluated the impact of the conditional likelihood and the marginal likelihood on the performance of different model comparison criteria using a simulation study. We would like to note that the simulation results showed a very salient pattern and were robust against the simulation settings. Our study echoed the previous literature and emphasized the use of marginal likelihood for the calculation of Bayesian model selection criteria when models contain latent variables.

We want to point out that when data are normally distributed, the marginal likelihood is recommended to use and DIC, WAIC, and LOO-CV calculated based on the marginal likelihood can almost guarantee the correct class enumeration. However, the performance of the model selection criteria based on different likelihoods were not systematically evaluated when data are contaminated by outliers. Previous research (e.g., Kim et al., 2021a) suggested the application of robust methods for dealing with the nonnormality in class enumeration in GMM. We expect that combining robust methods with marginal likelihood based model selection criteria may improve the model selection accuracy. Future research needs to be conducted towards this direction.

We also would like to note that in our study, the normality assumption is applied to the growth mixture model, with which a close form of the marginal likelihood is available. When a close form of the marginal likelihood cannot be obtained (e.g., a robust model using Student's t distributions), we need to numerically integrate the conditional likelihood with respect to the latent variables. Since numerical integration takes time, the entire class enumeration procedure may be slowed down. It is worth investigating ways to solve numerical integrations faster.



Acknowledgments This paper is based upon work supported by the National Science Foundation under grant no. SES-1951038. Correspondence concerning this article should be addressed to Xin Tong, Department of Psychology, University of Virginia. Email: xt8b@virginia.edu.

References

- Bauer, D. J. & Curran, P. J. (2003). Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. *Psychological Methods*, 8(3), 338–363. <https://doi.org/10.1037/1082-989X.8.3.338>
- Celeux, G., Forbes, F., Robert, C. P., & Titterton, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, 1, 651–673. <https://doi.org/10.1214/06-ba122>
- Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods*, 18, 186–219. <https://doi.org/10.1037/a0031609>
- Depaoli, S. (2014). The impact of inaccurate informative priors for growth parameters in Bayesian growth mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(2), 239–252. <https://doi.org/10.1080/10705511.2014.882686>
- Frankfurt, S., Frazier, P., Syed, M., & Jung, K. R. (2016). Using group-based trajectory and growth mixture modeling to identify classes of change trajectories. *The Counseling Psychologist*, 44(5), 622–660. <https://doi.org/10.1177/0011000016658097>
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Boca Raton, FL: Chapman and Hall/CRC.
- Hipp, J. R. & Bauer, D. J. (2006). Local solutions in the estimation of growth mixture models. *Psychological Methods*, 11(1), 36–53. <https://doi.org/10.1037/1082-989X.11.1.36>
- Kim, S., Tong, X., & Ke, Z. (2021a). Exploring class enumeration in Bayesian growth mixture modeling based on conditional medians. *Frontiers in Education*. A special research topic on Advances in Mixture Modeling. <https://doi.org/10.3389/educ.2021.624149>
- Kim, S., Tong, X., Zhou, J., & Boichuk, J. P. (2021b). Conditional median based Bayesian growth mixture modeling for nonnormal data. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-021-01655-w>
- Lu, Z., Zhang, Z., & Lubke, G. (2011). Bayesian inference for growth mixture models with latent class dependent missing data. *Multivariate Behavioral Research*, 46, 567–597.
- McDermott, P. A., Rovine, M. J., Reyes, R. S., Chao, J. L., Scruggs, R., Buek, K., & Fantuzzo, J. W. (2018). Trajectories of early education learning behaviors among children at risk: A growth mixture modeling approach. *Psychology in the Schools*, 55(10), 1205–1223. <https://doi.org/10.1002/pits.22145>
- Merkle, E. C., Furr, D., & Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: conditional versus marginal likelihoods. *Psychometrika*, 84(3), 802–829. <https://doi.org/10.1007/s11336-019-09679-0>
- Plummer, M. (2017). *Jags version 4.3.0 user manual*.
- Ram, N. & Grimm, K. J. (2009). Growth mixture modeling: a method for identifying differences in longitudinal change among unobserved groups. *International Journal of Behavioral Development*, 33(6), 565–576. <https://doi.org/10.1177/0165025409343765>
- Ren, L., Tong, X., Xu, W., Wu, Z., Zhou, X., & Hu, B. Y. (2021). Distinct patterns of organized activity participation and their associations with school readiness among Chinese preschoolers. *Journal of School Psychology*, 86, 100–119. <https://doi.org/10.1016/j.jsp.2021.03.007>
- Smith, K. V. & Ehlers, A. (2020). Cognitive predictors of grief trajectories in the first months of loss: A latent growth mixture model. *Journal of Consulting and Clinical Psychology*, 88(2), 93–105. <https://doi.org/10.1037/ccp0000438>
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. v. d. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583–639.
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Journal of Statistics and Computing*, 27, 1413–1432.
- Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11, 3571–3594.

A Bi-level Individualized Adaptive Learning Recommendation System Based on Topic Modeling



Jiawei Xiong, Jordan M. Wheeler , Hye-Jeong Choi, and Allan S. Cohen 

Abstract Adaptive learning offers real attention to individual students' differences and fits different needs from students. This study proposes a bi-level recommendation system with topic models, gradient descent, and a content-based filtering algorithm. In the first level, the learning materials were analyzed by a topic model, and topic proportions to each short item in each learning material were yielded as representation features. The second level contains a measurement component and a recommendation strategy component which employ gradient descent and content-based filtering algorithm to analyze personal profile vectors and make an individualized recommendation. An empirical data consists of cumulative assessments that were used as a demonstration of the recommendation process. Results have suggested that the distribution to the estimated values in the person profile vectors were related to the ability estimation from the Rasch model, and students with similar profile vectors could be recommended with the same learning material.

Keywords Individualized learning · Recommendation system · Topic model

1 Introduction

In recent years, especially during the pandemic, efforts have been made to expand online learning beyond the traditional classroom environment as it enables individuals to benefit from rich and high-quality learning resources (Dhawan, 2020; Liang & Hainan, 2019). The advantages of online learning have been recognized since it offers real attention to the individual differences and fits for different

J. Xiong (✉) · J. M. Wheeler · A. S. Cohen
University of Georgia, Athens, GA, USA
e-mail: jiawei.xiong@uga.edu

H.-J. Choi
The University of Georgia, Athens, GA, USA

The Human Resources Research Organization, Louisville, KY, USA

needs of students (Imhof et al., 2020). More importantly, it makes it possible to analyze students' latent information, here referred to as profile information or profile, through human-computer interactions, and in particular, with the maturity of cutting-edge learning analytics, individualized adaptive learning provides students the prospects of access to tailored learning instructions, guidance, and content (Mavroudi et al., 2018). With the popularization of remote education and online learning, offering individualized and adaptive learning resources is an emerging research topic (Cheng & Bu, 2020). Individualized adaptive learning systems aim to provide learning materials fit to the current status of a student, and the pace and of learning and instruction approach are optimized for the need of each student (United States Department of Education, 2017). The individualized adaptive learning system provides flexible adaptation beyond what can be accomplished in traditional classroom settings in terms of learning resources (Koedinger et al., 2013).

The purpose of adaptive learning is realized by using a recommendation system, which may recommend the next learning materials based on the psychometric results and possibly other individual-level characteristics (Chen et al., 2018). Specifically, the recommendation system requires three components, an information learning component, a measurement component, and a recommendation strategy component. The information learning component employs a learning model to analyze features from the learning materials such that each learning material's features can be represented in a numerical space. The features can be used as representations of a series of skills or attributes that are available in the learning system (Chen et al., 2018). Traditional recommendation systems suggest online learning materials based on students' interests, knowledge, and data from other students with similar interests (Romero et al., 2007). These traditional methods, which utilize vector space models (Castells et al., 2006) in the information learning component, have disadvantages in both effectiveness and scalability (Kuang et al., 2011). In addition, with a heterogeneous student population and many learning materials, the learning model can be complex, and thus calibrating the model requires expensive computation (Tang et al., 2019). Topic modeling such as the Latent Dirichlet allocation (LDA; Blei et al., 2003), a hierarchical Bayesian topic model, was used to obtain a low dimensional vector that denotes each online learning activity in various adaptive learning scenarios, such as online course recommendations (Lin et al., 2021) and online documents recommendation (Kuang et al., 2011). Compared with traditional recommendation systems based on the student-item interactions and similarity, the topic model-based recommendation systems can consider the learning portfolios' textual features (Cao et al., 2019). With the learned features, the measurement component can find the profile vectors for students which may reveal students' proficiency on each attribute (Chen et al., 2018). Given the features and profiles, the prediction component uses a predicting model that predicts the outcomes of a student studying under a particular set of materials, and sequentially makes recommendations to each student on what to learn at the next step, based on the current information it obtained from the aforementioned two components. Content-based filtering (CB; Ghauth & Abdullah, 2010) algorithm, using information about students and/or learning materials, has

been used as a prediction component in many recommendation systems (Bian & Xie, 2010; Romero et al., 2007). The CB focuses on the properties of learning materials, and learning materials' similarity is determined by the similarity in their features.

The cumulative assessment portfolio has been used as learning material in online learning (Pfennig, 2020). Cumulative assessments are widely used in many circumstances to determine at a particular time what students know and do not know and can help students get access to their learning achievements (Beagley & Capaldi, 2016; Ryan & Nykamp, 2000). The cumulative assessments are comprehensive and pre-assembled tests that assess students' knowledge of information from several didactic domains, and in which each assessment covers all previous contents. By using the cumulative assessments, instructors can identify a wide range of knowledge, skills, and concepts that students have mastered or not, so appropriate adjustments can be made to instructional practices and strategies toward the overall end-of-year expectations (den Boer et al., 2021). For example, some English language and arts (Georgia Center for Assessment, 2018) cumulative assessments were designed to collect evidence on student learning status, and serve as formative tools that can provide information on how well students understand concepts and their ability to demonstrate knowledge and skills in a particular content area or domain. In higher education, it is effective to intersperse several cumulative assessments throughout a course and the combined score on the assessments weighs in for the final course grade (den Boer et al., 2021). For example, the United States Medical Licensing Exam Step I assesses whether the examinees can successfully apply the knowledge of key concepts in basic sciences and is usually taken by medical school examinees at the end of the second year (USMLE, 2014). Some medical schools ask students to take the cumulative licensing examination before initiating clinical experiences (Cleghorn, 1986; Ryan & Nykamp, 2000). Given the fact that cumulative assessments have wide applications, this study selects a set of pre-assembled cumulative assessments as learning materials.

It is suggested that a good recommendation system should make full use of the information from both the students and the learning materials (Tang et al., 2019). Therefore, this study designs a bi-level structure. In the first level, the learning materials (i.e., cumulative assessments) were analyzed by a topic model and the topic proportions to each item stem in the cumulative assessment were yielded as representation features to the cumulative assessment. Although most educational applications with topic models adopt the LDA as a useful model (Wheeler et al., 2021; Xiong et al., 2019), the use of LDA as a topic model tool is useful for long documents such as the course syllabus (Apaza et al., 2014), and it suffers from the severe data sparsity in short text documents (Yan et al., 2013). For instance, the pre-assembled cumulative assessments may contain some short text items such as multiple-choice (MC) items, and which lengths are usually less than a passage or course syllabus content. Obviously, in such a circumstance, the use of LDA may cause sparse topic structures. To overcome the problem, in this first level, this study employs another topic model, called the bi-term topic model (BTM; Yan et al., 2013), which was designed to extract topic proportions for short

context, to analyze the learning materials (i.e., each cumulative assessment) and obtain each short item’s topic structure. The second level contains the measurement and recommendation strategy components which employ profile analysis and CB filtering algorithms. By proposing such a framework that applies both the BTM and CB filtering to recommend pre-assembled cumulative assessments with an empirical data demonstration, this recommendation system can analyze each student’s profile components based on their response scores to the completed assessments and then predict rating scores for new assessments. The empirical results suggested this design can recommend relevant assessments for each student, and realize individualized recommendations based on the bi-level framework.

2 Method

2.1 Bi-term Topic Model

BTM generates the bi-terms in the whole corpus to reveal topics by considering the word-pair relation. The bi-term here was referred to as an unordered word-pair co-occurred in a short context such as the example given in Table 1. The texts given in the Table 1 are all simple examples of short texts. After removing stopwords such as “I”, and stemming words into an original form such as changing from “apples” to “apple”, from “eating” to “eat”, the bi-terms were generated by construction word-pair combination in an unordered way.

The BTM graphical structure is represented in Fig. 1, and this generative process in the BTM can be described as:

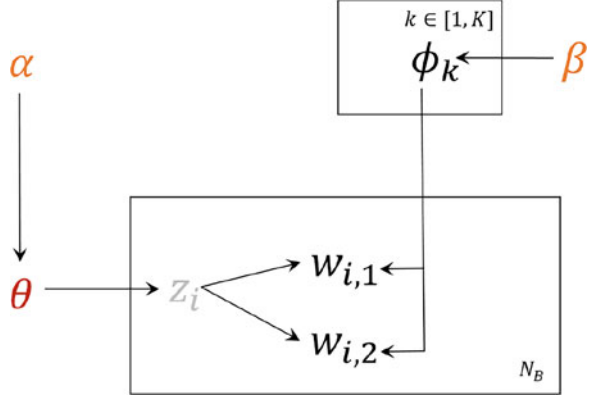
1. Draw a topic distribution θ from Dirichlet distribution with parameter α , i.e., $\theta \sim Dir(\alpha)$.
2. For each topic $k \in [1, \dots, K]$, draw a topic-specific word distribution ϕ_k from the Dirichlet distribution with parameter β , i.e., $\phi_k \sim Dir(\beta)$.
3. For each bi-term combination $b_i \in B$:
 - (a) draw a topic assignment $z_i \sim Multinomial(\theta)$
 - (b) draw two words, $w_{i,1}, w_{i,2} \sim Multinomial(\phi_{z_i})$

where N_B is the bi-term corpus which consists of all bi-terms given in a document collection, α is the prior distribution parameter to the topic distribution θ , β is the

Table 1 Simple bi-term examples

Text	Bi-terms
I visit an Apple store.	visit Apple, visit store, Apple store
I like eating apples.	like eat, like apple, eating apples
I love to watch Apple movies.	love watch, love Apple, love movie, watch Apple, watch movie, Apple movie

Fig. 1 BTM graphical structure



prior distribution parameter to the bi-term distribution ϕ_k . BTM directly models the word-co-occurrence pattern instead of a single word.

In this study, each pre-assembled cumulative assessment is treated as learning material, and adaptive learning happens after the completion of each assessment. Each MC item stem was modeled as a short text. Suppose the learning system consists of $i = 1, \dots, I$ cumulative assessments, while each of which consists of n_i MC items, then a total of $I \times n_i$ items are treated as a large collection of short documents. Suppose k topics were determined for the learning system, and then each item can be represented by a k -dimensional feature vector. Therefore, each cumulative assessment is represented by a $n_i \times k$ dimensional matrix. To determine the optimal number of topics k for the corpus, average Jensen-Shannon (JSD; Tong & Zhang, 2016) was used as criteria. JSD is a popular method of measuring the similarity between two probability distributions and is also known as a total divergence to the average. Given two discrete topic distributions T_s and T_v , the JSD is defined as Eq. 1.

$$JSD(T_s \| T_v) = \frac{1}{2} KLD \left(T_s \| \frac{T_s + T_v}{2} \right) + \frac{1}{2} KLD \left(T_v \| \frac{T_s + T_v}{2} \right) \quad (1)$$

where $KLD \left(T_s \| \frac{T_s + T_v}{2} \right)$ is the Kullback-Leibler divergence (KLD; Mei et al., 2007) of T_s from $\frac{T_s + T_v}{2}$, and the KLD is defined as in the Eq. 2.

$$KLD(P \| Q) = \sum_s P_s \log \frac{P_s}{Q_s} \quad (2)$$

The average JSD shown in Eq. 3 is used to calculate the average similarity among all topic distributions.

$$\overline{JSD} = \frac{\sum_{s,v} JSD(T_s \| T_v)}{k} \quad (3)$$

By applying JSD to the topic assignment for each item in the learning system, it will measure the distance and similarity between each document. The topic model with minimal average JSD is used as the optimal topic model. With these k topics, each MC item n 's features can be represented by a k -dimensional topic proportion vector $\mathbf{f}_n = (f_{n_1}, \dots, f_{n_k})$.

2.2 Loss Function and Gradient Descent

Suppose a random assessment i is given at the initial stage to a total of $j = 1, \dots, J$ students, while the remaining $(I - 1)$ assessments were waiting in the system to be sequentially recommended to students. With the k -dimensional feature vector $\mathbf{f}_{in} = (f_{in_1}, \dots, f_{in_k})$ for each MC item n in the assessment i , student j 's profile vector can also be defined as a k -dimensional vector $\boldsymbol{\alpha}_j = (\alpha_{j_1}, \dots, \alpha_{j_k})$. Each dimension to the profile vector serves as the weight or coefficient to items' feature vector. In addition, for each MC item in the assessment i , student j 's response can be scored as either correct or incorrect (i.e., $t_{jn} = 1/0$), so that cross-entropy (De Boer et al., 2005) is used as a loss function with t_{jn} serving as guiding labels, which is defined in the Eq. 4

$$L(\boldsymbol{\alpha}) = -\frac{1}{Jn_i} \left[\sum_{j=1}^J \left[\sum_{n=1}^{n_i} [t_{jn} \log(p_{jin}) + (1 - t_{jn}) \log(1 - p_{jin})] \right] \right] \quad (4)$$

where $p_{jin} = \sigma(\boldsymbol{\alpha}_i \mathbf{f}_{in}) = \sigma(\alpha_{j_1} f_{in_1} + \dots + \alpha_{j_k} f_{in_k})$, and the $\sigma(\cdot)$ represents a sigmoid function.

The gradient descent (Amos & Yarats, 2020) is used to minimize the loss function until it reaches convergence. The process of finding the minimized loss function is described in Table 2, where ρ is a positively defined learning rate, and $\nabla J(\boldsymbol{\alpha}_r)$ is the differential at $\boldsymbol{\alpha}_r$. The $\rho \nabla J(\boldsymbol{\alpha}_r)$ is subtracted from $\boldsymbol{\alpha}_r$ and moves toward the local minimum. So, a monotonic sequence $J(\boldsymbol{\alpha}_r) \geq J(\boldsymbol{\alpha}_{r+1}) \geq J(\boldsymbol{\alpha}_{r+2}) \geq \dots$ is obtained until convergence.

Table 2 Pseudo-code gradient descent algorithm

Algorithm	Gradient descent
For	$r = 1, 2, \dots$
	repeat
	until
	convergence
Output	$J(\boldsymbol{\alpha}_{r+1}) \& \boldsymbol{\alpha}_{r+1}$

3 Data and Analytic Framework

3.1 Data Description

Learning materials are used to construct a learning pool in which every learning material is pending to be recommended or not for each student. The learning pool in this study contains 8 science cumulative assessments (Georgia Center for Assessment, 2018) as learning materials, and each assessment contains 22 MC items. The assessments are designed to assess student learning on several sub-domains in science such as biology, physics, and chemistry. A link to the sample assessment was provided in the Appendix. Each MC item stem was pre-processed. This process includes stemming and lemmatization, and stop word removal. The stemming uses the stem of each word and cuts off the end or the beginning of the word such as the affixes of plural words. The lemmatization uses the context in which the word is being used and changes the word into the base forms such as the irregular verbs and irregular plural nouns. Stop words are high-frequency terms with little or no information and include words such as “the”, “and”, “is” etc. The cleaned item stems were treated as short texts and were modeled in BTM to extract representation vectors. The descriptive statistics to the length of clean item stems are listed in Table 3. The minimal length to the MC items is only 3, and the average length of these items is 10.530. Therefore, the lengths of items are relatively short and the use of BTM is appropriate.

Students’ response to each MC item was scored as either correct (1) or incorrect (0). Students’ responses t_{jn} to one learning material (i.e., one assessment containing 22 items) were used as guiding labels which are defined in Eq. 4. In this study, Assessment 4 was selected, and 492 students have responded to the 22 MC items in the assessment. All the response correctness t_{jn} given by these 492 students were used as the guiding labels to supervise the parameter estimation.

3.2 Bi-level Recommendation Framework

The bi-level recommendation system is shown in Fig. 2. The feature learning component in the first level employs the BTM described in Fig. 1 to extract feature matrix (dimension $n_i \times k$) for each assessment. The measurement component in the second level uses the J students’ responses to each item in one selected assessment and employs the gradient descent algorithm described in Table 2 to minimize the loss function to obtain J ’s k -dimensional vectors as students’ profile

Table 3 Descriptive statistics to MC items’ length in the learning pool

Min.	Mean	Max.	SD
3.000	10.530	32.001	10.812

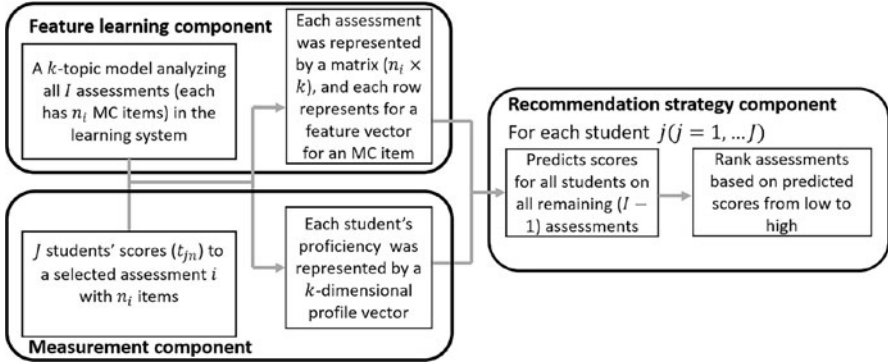


Fig. 2 Bi-level recommendation framework

vectors. The estimated profile vector was used as a quantified indicator of each student’s understanding of each of the k topics at this moment. Furthermore, with the k -dimensional student j ’s profile vector and remaining $I - 1$ assessments’ feature matrices, the recommendation strategy component predicts student j ’s score probability on each MC item in each remaining assessment with Eq. 5.

$$f(z_{jin}) = \frac{1}{1 + e^{-z_{jin}}} \tag{5}$$

where $z_{jin} = \alpha_j f_{in}$, and student j will be predicted to get a score $g_{jin} = 1$ for item n in assessment i when $f(z_{jin}) \geq 0.5$, and score $g_{jin} = 0$ when $f(z_{jin}) < 0.5$.

3.3 Analytic Procedures

The first step of this recommendation system uses the feature learning component to construct a representation matrix for each learning material. That is, by modeling every learning material i in the feature learning component, a k -column feature matrix was extracted by the BTM with JSD. In this study, each extracted feature matrix contains 22 rows and k columns, and each feature matrix serves as a representation matrix of the corresponding learning material. Once the representation matrices for all learning materials were constructed, then one learning material was randomly selected for all students and the remaining $I - 1$ learning materials are still in the learning pool. The second step of this recommendation system uses all students’ complete response patterns to estimate j th student’s k -dimensional profile vector $\alpha_j == (\alpha_{j1}, \dots, \alpha_{jk})$ in the measurement component with the loss function in Eq. 4 and gradient descent algorithm in Table 2. Finally, with the k -dimensional profile vectors for each student and remaining $I - 1$ assessments’ feature matrices, the recommendation strategy component predicts j th student’s

correctness probability on each MC item in each remaining assessment with Eq. 5 and predicts a score g_{jin} for each item n of j th student. The summation $g_{ji} = \sum_n g_{jin}$ will be used as j th student’s predicted total score on the assessment i in the learning system. After ranking the predicted total scores for all remaining $I - 1$ assessments from low to high, the system can make the next recommendation for each student.

4 Results

4.1 Step 1: Feature Learning Component with BTM and JSD

In the feature learning component, JSD was used as criteria for some exploratory topic models from 2 topics to 6 topics for all items in the learning system. Figure 3a is the average JSD values against the different number of topics, and the minimal JSD was achieved when the number of topics is four. So, it is suggested that the four-topic model fits best for all items in the learning system. After fitting a four-topic BTM, every item stem was characterized by a 4-dimensional vector and each dimension represents a topic’s proportion in the item stem. For example, the item n in assessment i can be characterized by a vector of $f_{in} = (0.1, 0.3, 0.1, 0.5)$ in which each value describes the topic distribution to this item such that 10% of words in the item belong to Topic 1 and 30% of words belong to Topic 2, etc. Table 4 lists the top 10 words under each of the four topics. Topic 1 can be described as words related to the chemistry process and natural resources, Topic 2 tends to employ words about the ecosystem, Topic 3 contains question words such as “select the class from the following samples”, and Topic 4 shows words from astronomy and physics.

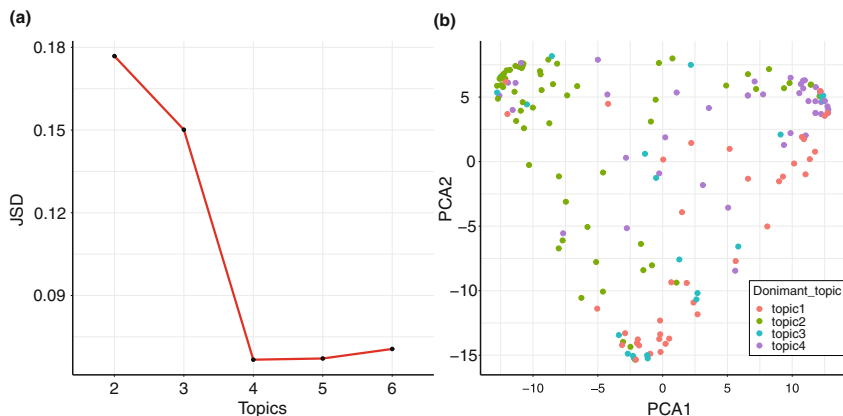


Fig. 3 (a) Average JSD against the number of topics; (b) PCA to the feature matrix

Table 4 Top 10 words from each of the four topics in the cumulative assessments

Topic 1		Topic 2		Topic 3		Topic 4	
water	0.034	model	0.019	class	0.021	earth	0.017
create	0.014	population	0.018	light	0.015	moon	0.015
rock	0.014	organism	0.016	give	0.015	sun	0.015
temperature	0.014	cell	0.016	form	0.012	weather	0.013
heat	0.013	base	0.015	see	0.011	layer	0.012
soil	0.011	show	0.014	sample	0.011	feather	0.011
hot	0.010	system	0.012	student	0.011	white	0.011
air	0.009	animal	0.012	leave	0.010	eye	0.010
plant	0.008	picture	0.010	need	0.009	gibbous	0.010
different	0.007	energy	0.009	question	0.009	move	0.010

Items may have different topic structures from each other; therefore, each item focuses on a domain combination in the cumulative assessment. If we denote the topic with the highest proportion as the dominant topic for an item, for instance, for the item with $f_{in} = (0.1, 0.3, 0.1, 0.5)$, its dominant topic was denoted as Topic 4. Figure 3b shows the principal component analysis (PCA; Chou & Wang, 2010) to the feature matrix with identity information from dominant topics. The two principles, PCA1 and PCA2, are associated with test domains. PCA1 represents environment-associated contents, and PCA2 stands for biology-associated contents. Each point represents an item in the two-dimensional space and each color represents a dominant topic. In this figure, items with similar topic distributions could be closer to each other, which indicates the items with similar topic distributions may measure similar test domains. We also noticed that some items with different dominant topics are mixed, which is because these items' topic distributions are flat such that the dominant topic has a close proportion to other topics.

The analysis in the feature learning component yielded 8 feature matrices and each has a dimension of 22×4 . In each feature matrix, every row represents an item feature vector f_{in} for n th item in i th assessment. Each dimension in the vector f_{in} represents for the topic proportion of n th item. Therefore, j th student's unknown profile vector is also 4-dimensional such that $\alpha_j = (\alpha_{j_1}, \alpha_{j_2}, \alpha_{j_3}, \alpha_{j_4})$.

4.2 Step 2: Measurement Component with Gradient Descent

After constructing the feature matrices for all learning materials, Assessment 4 was randomly selected for all 492 students. With the obtained feature matrix for Assessment 4, the response patterns were used to estimate j th student's profile vector $\alpha_j = (\alpha_{j_1}, \alpha_{j_2}, \alpha_{j_3}, \alpha_{j_4})$ in the measurement component. Students' profile vectors are obtained by gradient descent on the loss function. The gradient descent

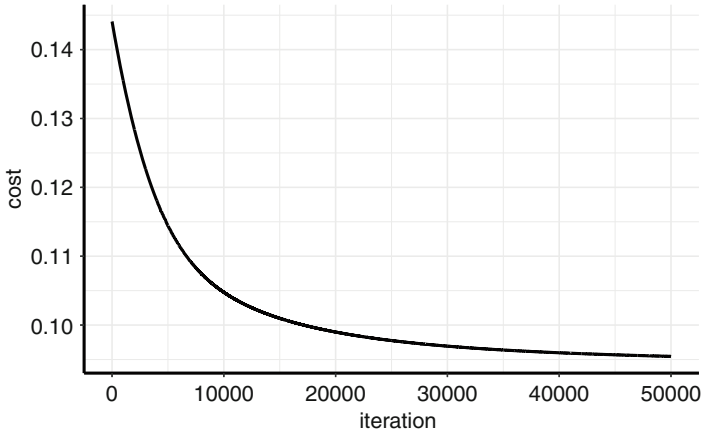


Fig. 4 Cost function against iterations

Table 5 Summary statistics on the estimated profile distribution

	Min.	Mean	Max.	SD
α_{j_1}	-1.045	-0.188	0.743	0.399
α_{j_2}	-1.039	0.201	1.498	0.364
α_{j_3}	-1.001	-0.017	0.859	0.339
α_{j_4}	-1.047	-0.084	0.686	0.364

changes are reflected in Fig. 4, which shows the cost decrease against each iteration. From Fig. 4, the estimation converged with increasing iterations finally and the cost was less than 0.1. The obtained estimations are student profile vectors that are coefficients to each dimension in the feature matrix.

The estimated profile vectors are summarized in Table 3. In Table 3, each row lists descriptive statistics of one dimension. As introduced, the estimated coefficient α_{jk} can be interpreted as j th student’s understanding of k th dimensional feature. For example, a student with the minimal value of $\alpha_{j_1} = -1.054$ may indicate that the student owns a relatively low understanding status of -1.054 on Topic 1, while a student with the maximal value of $\alpha_{j_1} = 0.743$ means that the student owns a relatively high understanding status to this topic. The standard deviations to the four dimensions were from 0.339 to 0.399. The α_{j_2} has the largest range from -1.039 to 1.498, while the other three dimensions are distributed between -1 and 1 (Table 5).

An item response analysis was conducted to help interpret profile vectors. Students’ correctness responses were analyzed by a Rasch model, which can calibrate students’ ability levels into logit scale and rank the logits on a one-dimensional continuum (Engelhard, 2013), to explore the relationship between students’ profile vector and students’ latent ability. By assuming there is a unidimensional ability of students for answering these items correctly, the density of calibrated Rasch ability is plotted in Fig. 5a, which is approximately normally distributed with a mean of 0.000. The minimal student ability value is -1.942 and the maximal ability value is

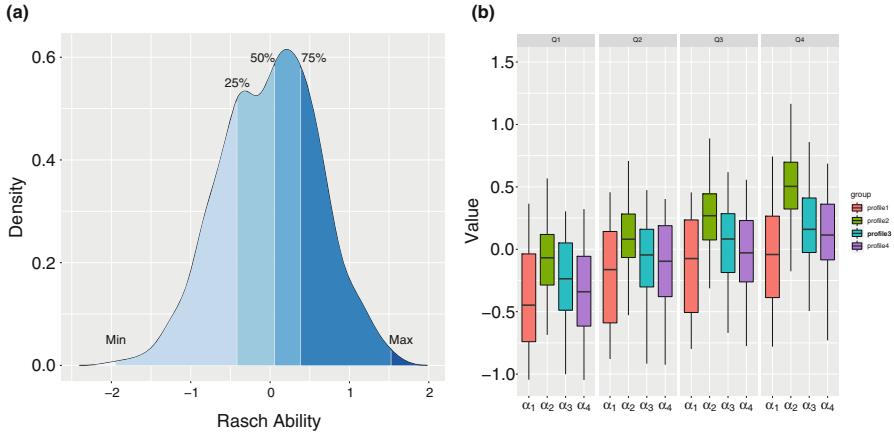


Fig. 5 (a) Rasch ability density with 4 quartiles; (b) Distribution of α in each quartile

1.521. The first quartile (the lowest 25%), second quartile (between 25.1 and 50%), and third quartile (50.1–75%) to the estimated Rasch ability are -0.413 , 0.056 , and 0.380 , respectively. By labeling students into four categories according to their Rasch ability such that Q1: $(-1.942, -0.413)$, Q2: $(-0.413, 0.056)$, Q3: $(0.056, 0.380)$, Q4: $(0.380, 1.521)$, the order from Q1 to Q4 also represents both the ability levels and the probability of answering items correctly are increasing.

Figure 5b shows the distribution of α 's based on every quartile category. It is clear to observe that each dimensional α value shows a trend of increasing from lower category to higher category and the α_2 exhibits the most obvious increase trend. Possibly Topic 2, containing words about the ecosystem, covers the main test sub-domain in the cumulative learning material. From Q1 to Q4, the Rasch estimates are increasing, and the probability of answering items correctly is increasing. Since each student's profile vector indicates the student's understanding status of certain topics, larger α will lead to a higher probability of taking each item correctly. This further verifies the homogeneity of Rasch analysis and profile analysis in terms of item correctness probability, and students with higher ability values are also likely to have higher profile values on each dimension.

4.3 Step 3: Recommendation Strategy Component with Predicted Total Scores

Given students' current understanding status of certain topics, their predicted total scores for all remaining assessments in the learning pool were calculated based on the profile vectors and learning matrices in the recommendation strategy component. Every two students tend to have different profile vectors unless they have the same

responding pattern to the 22 MC items. Therefore, the predicted patterns on each of the remaining learning materials for every two students could be different.

All 492 profile vectors were multiplied to each of the learning matrices of the remaining assessments for predicting j th student's n th item score g_{jin} of learning material i using Eq. 5. For j th student, the summation of all item scores within i th assessment $g_{ji} = \sum_n g_{jin}$ is used as the predicted total score on the assessment i in the learning system. These scores indicate a student's predicted achievements on the remaining learning materials that the student may obtain based on their current α values. Learning materials with lower scores indicate that the student may perform comparatively worse on that materials than the ones with higher scores. For each student, the predicted scores on remaining learning materials were ranked from low to high, and the learning material with the lowest score was recommended to the student for next-step learning. Figure 6 shows students' predicted scores distribution for each of the remaining learning materials, where the vertical axis stands for the predicted scores. In this figure, the predicted scores for each assessment range from 0 to 22. The predicted scores of Assessment 1 have a relatively lower 1st quartile value, which indicates that more students were predicted to have a lower score on Assessment 1. The predicted scores of Assessment 7 have a higher 3rd quartile value than other assessments, which means that more students were predicted to have a higher score on Assessment 7.

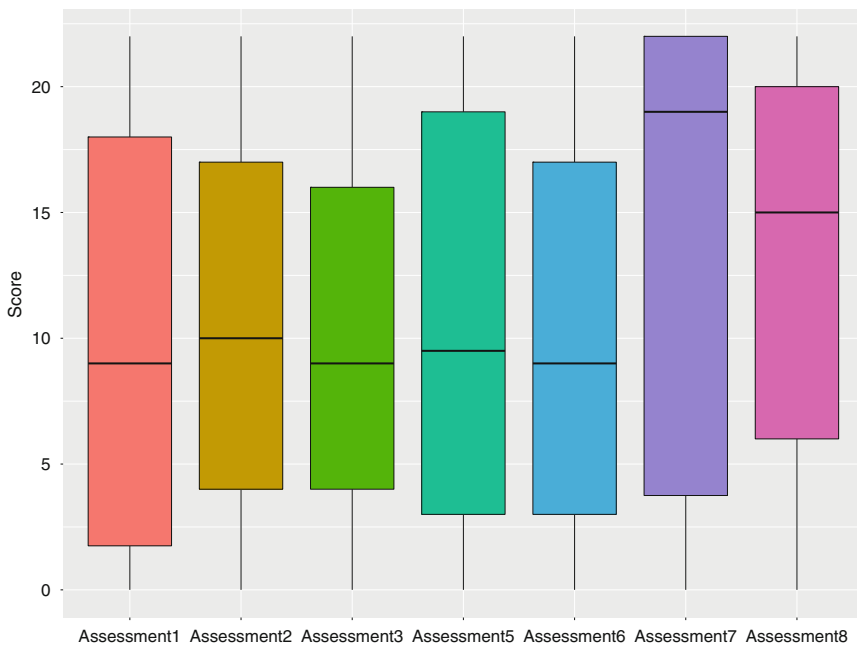


Fig. 6 Students' predicted scores distribution for each of the remaining learning materials

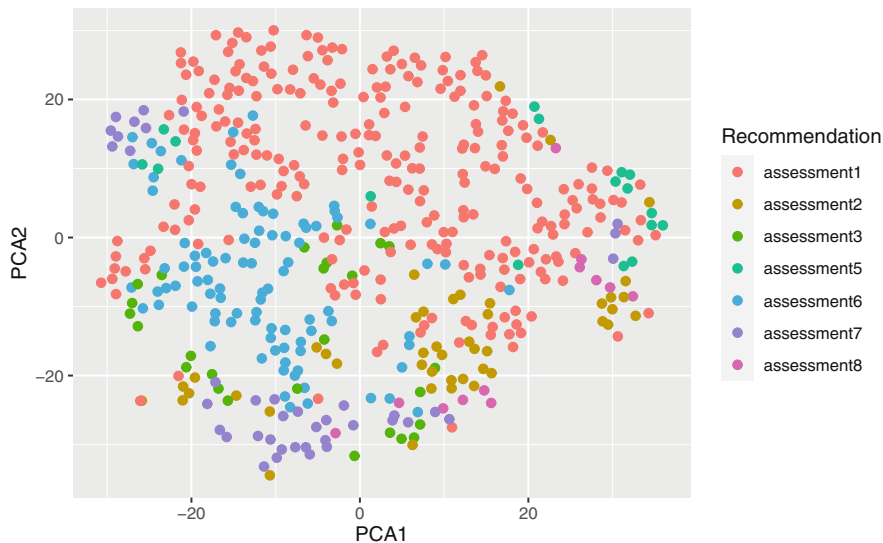


Fig. 7 PCA with recommended learning materials

Table 6 Correlations between each dimension of the profile vectors and each principal component

	PCA1	PCA2
α_{j_1}	0.524	0.229
α_{j_2}	0.711	0.575
α_{j_3}	0.190	0.207
α_{j_4}	0.250	0.544

Figure 7 shows the PCA on students’ profile vectors with colors indicating their recommendation results on the 7 accumulative assessments. In this figure, each point represents a student, and students with similar profile vectors were close to each other on this lower-dimensional space. It can be seen that many students were recommended to take Assessment 1, and which is consistent with the observation from Fig. 6 that more students were predicted to have a lower score on Assessment 1. In addition, students with similar profile vectors could be recommended with the same learning material.

Table 6 listed the correlations between each dimension of the profile vectors and each principal component, and correlations higher than 0.500 were in bold. PCA1 is strongly correlated with two dimensions of the profile vectors and PCA1 increases with increasing α_{j_1} and α_{j_2} . This component can be viewed as a measure of the values to α_{j_1} and α_{j_2} . Furthermore, we see that the first principal component correlates most strongly with α_{j_2} . Considering the interpretation of α_{j_1} and α_{j_2} , PCA1 may indicate students’ understanding status of concepts with the ecosystem and natural resources. PCA2 also correlates with two dimensions of the profile vectors, α_{j_2} and α_{j_4} . Similarly, this component can be viewed as a measure of

the understanding of Topic 2 and Topic 4, so PCA2 primarily stands for students' understanding status of concepts with astronomy and physics.

5 Discussion and Conclusion

This study proposed a bi-level recommendation system consisting of three components. An empirical study shows that, by employing the topic model and gradient descent algorithm, students profile vectors can be extracted and individualized recommendations can be made based on the predicted scores for the learning materials. The analysis also suggested that the distribution to the estimated values in the person profile vectors were related to the ability estimation from the Rasch model. Future researches can focus on a simulation study that explores the recovery accuracy of the profile vectors.

Although the recommendation component shows interesting findings and predicts individual scores on each item in the learning material, one thing still needs to note is that, for each student, learning materials with different score patterns may be predicted with the same final scores. This is because the sum scores for each learning material were used. For example, suppose a four-item cumulative assessment of predicted score patterns of $s = (0, 1, 0, 1)$ has the same sum score as another assessment with pattern $s = (1, 0, 1, 0)$, then both assessments could be recommended next. In addition, this empirical data demonstration used assessments that have the same length. However, when learning materials consist of learning materials with different item numbers, a biased situation may be produced as short-length learning materials may be preferred if the sum score is still used. To better process situations with these problems, one possible solution is that the recommendation component design in the future study could assess the psychometric properties that each item has such as the item difficulties, and items with different difficulties can be assigned different weights when sum score was used.

Appendix

Data

Cumulative Assessments are aligned and assess a representation of the Georgia Standards of Excellence (GSE). These cumulative forms can help teachers to gather strong evidence on student learning toward the overall end-of-year expectations at each grade level. The sample assessment items were provided on this web site: <https://www.lennconnections.com/assesslets-science>

R Code

```
#####
#####read data#####
#####
data<-read.csv(file='data.csv', header=T, sep=",", fill=T,
  stringsAsFactors = F)
#processing
data2 <- udpipe(data, "english")
biterns <- as.data.table(data2)[, cooccurrence(x = lemma,
  relevant = upos %in%
  c("NOUN",
  "ADJ", "VERB") &
  nchar(lemma) > 2 & !lemma
  %in%
  stopwords("en"),
  skipgram = 3),
  by = list(doc_id)]
data3 <- data2[, c("doc_id", "lemma")]

#####
#####decide optimal numbers#####
#####
cd_k<-seq(2,10)
#JSD
model=NULL
for (i in cd_k) {
  model[[i]] <- BTM(data3, biterns = biterns,
  k = i,
  alpha = 1,
  beta = 1,
  window = 3,
  iter = 5000, background = F,
  trace = F,detailed = F)
}

# Compute Jensen-Shannon Divergence for each value in model
scores <- predict(model[[1]], newdata = data3)
colnames(scores)<-c("topic1","topic2","topic3","topic4")
JSD <- function(p, q) {
m <- 0.5 * (p + q)
divergence <-
  0.5 * (sum(p * log(p / m)) + sum(q * log(q / m)))
return(divergence)
}

n <- dim(scores)[1]
X <- matrix(rep(0, n*n), nrow=n, ncol=n)
indexes <- t(combn(1:nrow(scores), m=2))
for (r in 1:nrow(indexes)) {
i <- indexes[r, ][1]
j <- indexes[r, ][2]
```

```

p <- scores[i, ]
q <- scores[j, ]
X[i, j] <- JSD(p,q)
}

#####
#####Estimation and predict#####
#####
#####read students' response data
student = read.csv(file='4_Cumulative_Assesslet.csv',
                    header=T, sep=",",
                    fill=T,stringsAsFactors = F)
#M4 is feature matrix of 4th assessment
M4 = read.csv(file='M4.csv', header=T, sep=",",
              fill=T,stringsAsFactors = F)

X = as.matrix(M4)
y= as.matrix(student)
N= dim(y) [1]*dim(y) [2]
theta.init = matrix(rnorm(n=dim(X) [2]*dim(y) [1],
                          mean=0,sd = 1),
                    nrow=dim(y) [1],ncol=dim(X) [2], byrow=T)
e = y - theta.init%*%t(X)
grad.init = -(2/N)*(e)%*%X
theta = theta.init - eta*(1/N)*grad.init
l2loss = c()
for(i in 1:iters){
myMatrix = y - theta%*%t(X)
# empty matrix for the results
squaredMatrix = matrix(nrow=dim(myMatrix) [1],
                       ncol=dim(myMatrix) [2])

for(i in 1:nrow(myMatrix)) {
for(j in 1:ncol(myMatrix)) {
squaredMatrix[i,j] = myMatrix[i,j]^2
}
}
l2loss = c(l2loss,sqrt(sum(squaredMatrix)))
e = y - theta%*%t(X)
grad = -(2/N)*e%*%X
theta = theta - eta*(2/N)*grad
# empty matrix for the results
squaredMatrix2 = matrix(nrow=dim(grad) [1],
                       ncol=dim(grad) [2])

for(i in 1:nrow(grad)) {
for(j in 1:ncol(grad)) {
squaredMatrix2[i,j] = grad[i,j]^2
}
}
if(sqrt(sum(squaredMatrix2)) <= epsilon){
break
}
}
}

```

```

values<-list("coef" = theta, "l2loss" = l2loss)

h=sigmoid(X%*%t(theta.init))
sum(diag(-y%*%log(h) - (1-y)%*%log(1-h)))/m
#sigmoid function, inverse of logit
sigmoid <- function(z){1/(1+exp(-z))}

#initialize theta
theta <- matrix(rnorm(n=dim(X)[2]*dim(y)[1],
                    mean=0,sd = 1),
               nrow=dim(y)[1],ncol=dim(X)[2], byrow=T)
#comput GD
compCost<-function(para){
m <- dim(y)[1]*dim(y)[2]
j=0
for (i in seq(1,492*4,by=4)) {
k=match(i,seq(1,492*4,by=4))
l1_1=sigmoid(colSums(para[i:(i+3)]*t(X)))
l1 <- log(l1_1)
l2 <- log(1-l1_1)
j=j+sum(y[k,]*l1+(1-y[k,])*l2)
}
J=-j/m
}

```

References

- Amos, B., & Yarats, D. (2020). *The differentiable cross-entropy method*. Paper presented at the International Conference on Machine Learning.
- Apaza, R. G., Cervantes, E. V., Quispe, L. C., & Luna, J. O. (2014). *Online courses recommendation based on LDA*. Paper presented at the SIMBig.
- Beagley, J. E., & Capaldi, M. (2016). The effect of cumulative tests on the final exam. *Primus*, 26(9), 878–888.
- Bian, L., & Xie, Y. (2010). *Research on the adaptive strategy of adaptive learning system*. Paper presented at the International Conference on Technologies for E-Learning and Digital Entertainment.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993–1022.
- Cao, Y., Li, W., & Zheng, D. (2019). A hybrid recommendation approach using LDA and probabilistic matrix factorization. *Cluster Computing*, 22(4), 8811–8821.
- Castells, P., Fernandez, M., & Vallet, D. (2006). An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Transactions on Knowledge Data Engineering*, 19(2), 261–272.
- Chen, Y., Li, X., Liu, J., & Ying, Z. (2018). Recommendation system for adaptive learning. *Applied Psychological Measurement*, 42(1), 24–41.
- Cheng, Y., & Bu, X. (2020). *Research on key technologies of personalized education resource recommendation system based on big data environment*. Paper presented at the Journal of Physics: Conference Series.

- Chou, Y.-T., & Wang, W.-C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals. *Educational and Psychological Measurement, 70*(5), 717–731.
- Cleghorn, G. D. (1986). Policies of US medical schools on the use of the NBME Part I and Part II examinations. *Journal of Medical Education, 61*(12), 954–957.
- De Boer, P.-T., Kroese, D. P., Mannor, S., & Rubinstein, R. (2005). A tutorial on the cross-entropy method. *Annals of Operations Research, 134*(1), 19–67.
- den Boer, A. W., Verkoeijen, P. P., & Heijltjes, A. E. (2021). Comparing formative and summative cumulative assessment: Two field experiments in an applied university engineering course. *Psychology Learning & Teaching, 20*(1), 128–143.
- Dhawan, S. (2020). Online learning: A panacea in the time of COVID-19 crisis. *Journal of Educational Technology Systems, 49*(1), 5–22.
- Engelhard, G. Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.
- Georgia Center for Assessment. (2018). *English language and Arts cumulative Assessment*.
- Ghauth, K. I., & Abdullah, N. A. (2010). Learning materials recommendation using good learners' ratings and content-based filtering. *Educational Technology Research and Development, 58*(6), 711–727.
- Imhof, C., Bergamin, P., & McGarrity, S. (2020). Implementation of adaptive learning systems: Current state and potential. In *Online teaching and learning in higher education* (pp. 93–115). Springer.
- Koedinger, K. R., Brunskill, E., Baker, R. S., McLaughlin, E. A., & Stamper, J. (2013). New potentials for data-driven intelligent tutoring system development and optimization. *AI Magazine, 34*(3), 27–41.
- Kuang, W., Luo, N., & Sun, Z. (2011). *Resource recommendation based on topic model for educational system*. Paper presented at the 2011 6th IEEE Joint International Information Technology and Artificial Intelligence Conference.
- Liang, Q., & Hainan, N. C. (2019). *Adaptive learning model and implementation based on big data*. Paper presented at the 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD).
- Lin, Q., He, S., & Deng, Y. (2021). Method of personalized educational resource recommendation based on LDA and learner's behavior. *The International Journal of Electrical Engineering & Education, 0020720920983511*.
- Mavroudi, A., Giannakos, M., & Krogstie, J. (2018). Supporting adaptive learning pathways through the use of learning analytics: Developments, challenges and future opportunities. *Interactive Learning Environments, 26*(2), 206–220.
- Mei, Q., Shen, X., & Zhai, C. (2007). *Automatic labeling of multinomial topic models*. Paper presented at the Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- Pfennig, A. (2020). *Improving learning outcome for GSL (German as a Second Language) students in a blended learning cumulative assessment material science course*. Paper presented at the Int. Conf. on Education and E-Learning ICEEL 2020.
- Romero, C., Ventura, S., Delgado, J. A., & De Bra, P. (2007). *Personalized links recommendation based on data mining in adaptive educational hypermedia systems*. Paper presented at the European conference on technology enhanced learning.
- Ryan, G. J., & Nykamp, D. (2000). Use of cumulative examinations at US schools of pharmacy. *American Journal of Pharmaceutical Education, 64*(4), 409–412.
- Tang, X., Chen, Y., Li, X., Liu, J., & Ying, Z. (2019). A reinforcement learning approach to personalized learning recommendation systems. *British Journal of Mathematical and Statistical Psychology, 72*(1), 108–135.
- Tong, Z., & Zhang, H. (2016). *A text mining research based on LDA topic modelling*. Paper presented at the International Conference on Computer Science, Engineering and Information Technology.

- United States Department of Education. (2017). *Reimagining the role of technology in education: 2017 National Education Technology Plan update*. Washington, DC Retrieved from <https://tech.ed.gov/files/2017/01/NETP17.pdf>
- USMLE. (2014). Federation of State Medical Boards of the United States and the National Board of Medical Examiners. *USMLE Bulletin of Information*.
- Wheeler, J. M., Cohen, A. S., Xiong, J., Lee, J., & Choi, H.-J. (2021). Sample size for latent Dirichlet allocation of constructed-response items. In *Quantitative Psychology* (pp. 263–273). Springer.
- Xiong, J., Choi, H.-J., Kim, S., Kwak, M., & Cohen, A. S. (2019). Topic Modeling of Constructed-Response Answers on Social Study Assessments. In *The annual meeting of the psychometric Society* (pp. 263–274). Springer, Cham.
- Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). *A biterm topic model for short texts*. Paper presented at the Proceedings of the 22nd International Conference on World Wide Web.

Derivation of the Percentile Based Parameters for Tukey HH, HR and HQ Distributions



Yevgeniy Ptukhin, Yanyan Sheng, and Todd Headrick

Abstract In the statistical literature, there exist several systems of distributions and types of transformations. These systems can be estimated using the method of moments (MOM), the method of L-moments (MOL), or the method of percentiles (MOP). The MOM has been widely used for deriving conventional moment-based estimators of different parameters. Although the method is comparatively simple and produces consistent estimators, these estimators can be biased, affected by outliers, or can have large variances. The MOP provides a useful alternative to the MOM when the distributions are non-normal, specifically being more computationally efficient in terms of estimating population parameters. In this paper, we focus on the Tukey HH, HR and HQ distributions, as extensions of the Tukey g-h (GH) family, to theoretically derive the parameters based on the use of the MOP. More specifically, closed-form solutions are obtained for the fifth-ordered percentile based system parameters of these distributions.

Keywords Tukey distribution · Power method · Method of Percentiles

1 Introduction

A few types of transformations and systems of distributions are present in the statistical literature. They include the Pearson system (Pearson, 1895, 1901, 1916), the Burr system (Burr, 1942, 1973; Burr & Cislak, 1968), the Johnson system

Y. Ptukhin (✉)

Department of Mathematics, Statistics and Computer Science, Monmouth College, Monmouth, USA

e-mail: yptukhin@monmouthcollege.edu

Y. Sheng

University of Chicago, Chicago, IL, USA

T. Headrick

Southern Illinois University, Carbondale, IL, USA

(Johnson, 1949), the power method (Fleishman, 1978; Headrick, 2002), and the Tukey g-h (GH) family with its extensions (Hoaglin, 1985; Morgenthaler & Tukey, 2000; Tukey, 1960, 1977). The estimation methods for these systems consist of the method of moments (MOM, Headrick et al., 2008; Kowalchuk & Headrick, 2010), the method of L-moments (MOL, Headrick & Pant, 2012), and the method of percentiles (MOP, Kuo & Headrick, 2014).

The MOM has been widely used for deriving conventional moment-based estimators of different parameters. For instance, its use has been well established with the power method (Headrick, 2010), Johnson system (Johnson, 1949), Burr distributions and their extensions (Burr, 1942, 1973; Yari & Tondpour, 2017). Furthermore, the MOM estimates have been developed for Tukey g-h family and its extensions of HH, HR and HQ distributions, where univariate moments up to the 4th moment (Headrick et al., 2008) as well as multivariate moments (Kowalchuk & Headrick, 2010) have been introduced.

Even though MOM is relatively simple and produces consistent estimators, these estimators can be biased, affected by outliers, or can have large variances (Headrick, 2010). The MOP offers a useful alternative to the MOM for situations where it is difficult or impossible to use the MOM for the distribution moments, e.g. for chi-square distributions (Kuo & Headrick, 2017). Indeed, Kuo and Headrick (2014, 2017) demonstrated that the fitting, estimation, relative bias, and relative error are superior for MOP-based characterization of Tukey g-and-h distributions compared with the MOM-based characterization.

The MOP estimation of the fifth-ordered percentile-based system of the Tukey g-h distribution was introduced by Kuo and Headrick (2017). Even though MOP estimators have been derived for Burr type III and type XII distributions (Pant & Headrick, 2017), and the power method (Kuo & Headrick, 2017), such estimates for Tukey HH, HQ and HR distributions, remain to be derived to complement existing MOM and MOL estimates for those distributions. In view of the above, this paper focuses on the MOP estimation by deriving in closed form the parameters for the fifth-ordered percentile based (Kuo & Headrick, 2017) system for the Tukey HH, HQ, and HR families of distributions.

2 Tukey g-h Family of Distributions and the Extensions

The family of symmetric H distributions (Tukey, 1960) was introduced by Tukey for the purpose of constructing nonnormal random deviates. This family uses the monotone transformation of standard normal random deviates (Z) with the following inverse cumulative distribution function:

$$F^{-1}(h) = Z * \exp\left(0.5hZ^2\right), \quad \text{where } h > 0, \quad (1)$$

which creates heavier tails than the normal probability density function.

The family of H distributions has shown its usefulness in different situations such as modeling stock returns (Badrinath & Chatterjee, 1988, 1991), financial times stock exchange index returns (Mills, 1995), returns of aluminum and zinc (Fischer et al., 2003), operational risk (Guegan & Hassani, 2009), and solar flare data (Goerg, 2011).

Further research adds more flexibility to the topic of H distributions. The inverse cumulative distribution functions of the g and g-h families were derived (Hoaglin, 1985; Tukey, 1977) and are as follows:

$$F^{-1}(g) = (\exp(gZ) - 1) / g, \quad h = 0 \text{ (lognormal)} \tag{2}$$

$$F^{-1}(gh) = ((\exp(gZ) - 1) / g) * \exp(0.5hZ^2), \quad g \neq 0, h > 0. \tag{3}$$

Still, in contrast to the Pearson (1895, 1901, 1916) system of distributions, the family of g-h monotonic distributions does not cover the entire set of values for the skewness and kurtosis (Tukey, 1977). To address this issue, additional families denoted as HH, HR, and HQ distributions were developed (Morgenthaler & Tukey, 2000) to serve as extensions of the Tukey family. Specifically, the HH distribution is an asymmetric generalization of the family of H distributions. In place of the one parameter of h, a pair of parameters (h_L and h_R – for dealing with left and right tails separately) is considered as

$$F^{-1}(h) = \begin{cases} Z * \exp(0.5h_L Z^2) & Z \leq 0 \\ Z * \exp(0.5h_R Z^2) & Z > 0 \end{cases} \quad h_L \neq h_R \tag{4}$$

for h_R ≥ 0 and h_L ≥ 0.

The HQ family of distributions was developed to expand the tail elongation, so the exponent was modified to encapsulate the additional term 0.25qZ⁴. The formula for the inverse cumulative distribution function of the HQ distribution is

$$F^{-1}(h, q) = Z * \exp(0.5hZ^2 + 0.25qZ^4) \tag{5}$$

for q ≥ 0, h ≥ 0 or h < 0, q ≥ h²/4.

The HR family of distributions features both heavy tails and shape affected. For this family, the formula for the inverse cumulative distribution function is given as

$$F^{-1}(h, r) = Z * \exp(hZ^2 / (2 + rZ^2)) \tag{6}$$

for r ≥ 0 and h > -2r.

3 Derivation of the Fifth-Ordered Percentile Based System Parameters

The percentiles (θ_p) associated with a Tukey family of distributions can be obtained by making use of the standard normal distribution, and hence the commonly used location, scale, and shape parameters are defined as (Karian & Dudewicz, 2011, p. 172–173)

$$\alpha_1 = \theta_{0.5}, \quad -\infty < \alpha_1 < \infty$$

$$\alpha_2 = \theta_{0.9} - \theta_{0.1}, \quad \alpha_2 \geq 0$$

$$\alpha_3 = \frac{\theta_{0.5} - \theta_{0.1}}{\theta_{0.9} - \theta_{0.5}}, \quad \alpha_3 \geq 0$$

$$\alpha_4 = \frac{\theta_{0.75} - \theta_{0.25}}{\theta_{0.9} - \theta_{0.1}}, \quad 0 \leq \alpha_4 \leq 1,$$

for the median, inter-decile range, left-right tail-weight ratio (related to skew), and tail-weight factor (related to kurtosis), respectively. More recently, Kuo and Headrick (2017) extended them to a more general fifth-ordered percentile based system:

$$\alpha_1 = \theta_{0.5}$$

$$\alpha_2 = \theta_{0.9} - \theta_{0.1}$$

$$\alpha_3 = \frac{\theta_{0.7} - \theta_{0.5}}{\theta_{0.5} - \theta_{0.3}}$$

$$\alpha_4 = \frac{\theta_{0.625} - \theta_{0.375}}{\theta_{0.7} - \theta_{0.3}}$$

$$\alpha_5 = \frac{\theta_{0.5} - \theta_{0.1}}{\theta_{0.9} - \theta_{0.5}}$$

$$\alpha_6 = \frac{\theta_{0.75} - \theta_{0.25}}{\theta_{0.9} - \theta_{0.1}}.$$

The derivation of the general percentile based system parameters for the HH, HQ, and HR distributions begins by substituting the standard normal distribution percentiles (Z_p) into the inverse cumulative function, and specifically for the HR family of distributions, the parameters $\alpha_3, \alpha_4, \alpha_5, \alpha_6$ are derived as follows

$$\begin{aligned} \alpha_3 &= \frac{F^{-1}(Z_{0.7}) - F^{-1}(Z_{0.5})}{F^{-1}(Z_{0.5}) - F^{-1}(Z_{0.3})} \\ &= \frac{(Z_{0.7}) \exp\left(hZ_{0.7}^2 / (2 + rZ_{0.7}^2)\right) - (Z_{0.5}) \exp\left(hZ_{0.5}^2 / (2 + rZ_{0.5}^2)\right)}{(Z_{0.5}) \exp\left(hZ_{0.5}^2 / (2 + rZ_{0.5}^2)\right) - (Z_{0.3}) \exp\left(hZ_{0.3}^2 / (2 + rZ_{0.3}^2)\right)} = 1 \end{aligned} \tag{7}$$

$$\begin{aligned} \alpha_4 &= \frac{F^{-1}(Z_{0.625}) - F^{-1}(Z_{0.375})}{F^{-1}(Z_{0.7}) - F^{-1}(Z_{0.3})} \\ &= \frac{(Z_{0.625}) \exp\left(hZ_{0.625}^2 / (2 + rZ_{0.625}^2)\right) - (Z_{0.375}) \exp\left(hZ_{0.375}^2 / (2 + rZ_{0.375}^2)\right)}{(Z_{0.7}) \exp\left(hZ_{0.7}^2 / (2 + rZ_{0.7}^2)\right) - (Z_{0.3}) \exp\left(hZ_{0.3}^2 / (2 + rZ_{0.3}^2)\right)} \end{aligned} \tag{8}$$

$$\begin{aligned} \alpha_5 &= \frac{F^{-1}(Z_{0.5}) - F^{-1}(Z_{0.1})}{F^{-1}(Z_{0.9}) - F^{-1}(Z_{0.5})} \\ &= \frac{(Z_{0.5}) \exp\left(hZ_{0.5}^2 / (2 + rZ_{0.5}^2)\right) - (Z_{0.1}) \exp\left(hZ_{0.1}^2 / (2 + rZ_{0.1}^2)\right)}{(Z_{0.9}) \exp\left(hZ_{0.9}^2 / (2 + rZ_{0.9}^2)\right) - (Z_{0.5}) \exp\left(hZ_{0.5}^2 / (2 + rZ_{0.5}^2)\right)} \\ &= \frac{- (Z_{0.1}) \exp\left(hZ_{0.1}^2 / (2 + rZ_{0.1}^2)\right)}{(Z_{0.9}) \exp\left(hZ_{0.9}^2 / (2 + rZ_{0.9}^2)\right)} = 1 \end{aligned} \tag{9}$$

$$\begin{aligned} \alpha_6 &= \frac{F^{-1}(Z_{0.75}) - F^{-1}(Z_{0.25})}{F^{-1}(Z_{0.9}) - F^{-1}(Z_{0.1})} \\ &= \frac{(Z_{0.75}) \exp\left(hZ_{0.75}^2 / (2 + rZ_{0.75}^2)\right) - (Z_{0.25}) \exp\left(hZ_{0.25}^2 / (2 + rZ_{0.25}^2)\right)}{(Z_{0.9}) \exp\left(hZ_{0.9}^2 / (2 + rZ_{0.9}^2)\right) - (Z_{0.1}) \exp\left(hZ_{0.1}^2 / (2 + rZ_{0.1}^2)\right)} \\ &= \frac{2 (Z_{0.75}) \exp\left(hZ_{0.75}^2 / (2 + rZ_{0.75}^2)\right)}{2 (Z_{0.9}) \exp\left(hZ_{0.9}^2 / (2 + rZ_{0.9}^2)\right)} = \frac{(Z_{0.75}) \exp\left(hZ_{0.75}^2 / (2 + rZ_{0.75}^2)\right)}{(Z_{0.9}) \exp\left(hZ_{0.9}^2 / (2 + rZ_{0.9}^2)\right)}. \end{aligned} \tag{10}$$

For the HQ distribution, the 5fth-ordered percentile based system parameters are derived as

$$\begin{aligned}
 \alpha_3 &= \frac{F^{-1}(Z_{0.7}) - F^{-1}(Z_{0.5})}{F^{-1}(Z_{0.5}) - F^{-1}(Z_{0.3})} \\
 &= \frac{(Z_{0.7}) \exp(0.5hZ_{0.7}^2 + 0.25qZ_{0.7}^4) - (Z_{0.5}) \exp(0.5hZ_{0.5}^2 + 0.25qZ_{0.5}^4)}{(Z_{0.5}) \exp(0.5hZ_{0.5}^2 + 0.25qZ_{0.5}^4) - (Z_{0.3}) \exp(0.5hZ_{0.3}^2 + 0.25qZ_{0.3}^4)} \\
 &= \frac{(Z_{0.7}) \exp(0.5hZ_{0.7}^2 + 0.25qZ_{0.7}^4)}{- (Z_{0.3}) \exp(0.5hZ_{0.3}^2 + 0.25qZ_{0.3}^4)} = 1
 \end{aligned} \tag{11}$$

$$\begin{aligned}
 \alpha_4 &= \frac{F^{-1}(Z_{0.625}) - F^{-1}(Z_{0.375})}{F^{-1}(Z_{0.7}) - F^{-1}(Z_{0.3})} \\
 &= \frac{(Z_{0.625}) \exp(0.5hZ_{0.625}^2 + 0.25qZ_{0.625}^4) - (Z_{0.375}) \exp(0.5hZ_{0.375}^2 + 0.25qZ_{0.375}^4)}{(Z_{0.7}) \exp(0.5hZ_{0.7}^2 + 0.25qZ_{0.7}^4) - (Z_{0.3}) \exp(0.5hZ_{0.3}^2 + 0.25qZ_{0.3}^4)}
 \end{aligned} \tag{12}$$

$$\begin{aligned}
 \alpha_5 &= \frac{F^{-1}(Z_{0.5}) - F^{-1}(Z_{0.1})}{F^{-1}(Z_{0.9}) - F^{-1}(Z_{0.5})} \\
 &= \frac{(Z_{0.5}) \exp(0.5hZ_{0.5}^2 + 0.25qZ_{0.5}^4) - (Z_{0.1}) \exp(0.5hZ_{0.1}^2 + 0.25qZ_{0.1}^4)}{(Z_{0.9}) \exp(0.5hZ_{0.9}^2 + 0.25qZ_{0.9}^4) - (Z_{0.5}) \exp(0.5hZ_{0.5}^2 + 0.25qZ_{0.5}^4)} \\
 &= \frac{- (Z_{0.1}) \exp(0.5hZ_{0.1}^2 + 0.25qZ_{0.1}^4)}{(Z_{0.9}) \exp(0.5hZ_{0.9}^2 + 0.25qZ_{0.9}^4)} = 1
 \end{aligned} \tag{13}$$

$$\begin{aligned}
 \alpha_6 &= \frac{F^{-1}(Z_{0.75}) - F^{-1}(Z_{0.25})}{F^{-1}(Z_{0.9}) - F^{-1}(Z_{0.1})} \\
 &= \frac{(Z_{0.75}) \exp\left(\frac{hZ_{0.75}^2}{2} + \frac{qZ_{0.75}^4}{4}\right) - (Z_{0.25}) \exp\left(\frac{hZ_{0.25}^2}{2} + \frac{qZ_{0.25}^4}{4}\right)}{(Z_{0.9}) \exp\left(\frac{hZ_{0.9}^2}{2} + \frac{qZ_{0.9}^4}{4}\right) - (Z_{0.1}) \exp\left(\frac{hZ_{0.1}^2}{2} + \frac{qZ_{0.1}^4}{4}\right)} \\
 &= (Z_{0.75}/Z_{0.9}) \exp\left(\left(\frac{h}{2}\right)(Z_{0.75}^2 - Z_{0.9}^2) - \left(\frac{q}{4}\right)(Z_{0.75}^4 - Z_{0.9}^4)\right).
 \end{aligned} \tag{14}$$

Since parameter q affects the tail elongation in HQ distributions, this parameter is convenient for approximating heavy-tailed distributions.

The derivation of the 5th-ordered percentile based system parameters for the HH family is based individually on h_L and h_R ($h_L \neq h_R$) and depends on the $(100p)$ th percentiles, $Z_{p(L)}$ and $Z_{p(R)}$ from the standard normal distribution for left and right tails, respectively:

$$\begin{aligned}\alpha_l^3 &= \frac{F^{-1}(Z_{0.7}) - F^{-1}(Z_{0.5(L)})}{F^{-1}(Z_{0.5(R)}) - F^{-1}(Z_{0.3})} \\ &= \frac{(Z_{0.7}) \exp(0.5h_L Z_{0.7}^2) - (Z_{0.5(L)}) \exp(0.5h_L Z_{0.5(L)}^2)}{(Z_{0.5(R)}) \exp(0.5h_R Z_{0.5(R)}^2) - (Z_{0.3}) \exp(0.5h_R Z_{0.3}^2)} \\ &= \frac{(Z_{0.7(L)}) \exp(0.5h_L Z_{0.7(L)}^2)}{(Z_{0.3(R)}) \exp(0.5h_R Z_{0.3(R)}^2)}\end{aligned}\quad (15)$$

$$\begin{aligned}\alpha_3^R &= \frac{F^{-1}(Z_{0.7}) - F^{-1}(Z_{0.5})}{F^{-1}(Z_{0.5}) - F^{-1}(Z_{0.3})} \\ &= \frac{(Z_{0.7}) \exp(0.5h_R Z_{0.7}^2) - (Z_{0.5(R)}) \exp(0.5h_R Z_{0.5(R)}^2)}{(Z_{0.5(L)}) \exp(0.5h_L Z_{0.5(L)}^2) - (Z_{0.3}) \exp(0.5h_L Z_{0.3}^2)} \\ &= \frac{(Z_{0.7(R)}) \exp(0.5h_R Z_{0.7(R)}^2)}{(Z_{0.3(L)}) \exp(0.5h_L Z_{0.3(L)}^2)}\end{aligned}\quad (16)$$

$$\begin{aligned}\alpha_l^4 &= \frac{F^{-1}(Z_{0.625}) - F^{-1}(Z_{0.375(L)})}{F^{-1}(Z_{0.7(R)}) - F^{-1}(Z_{0.3})} \\ &= \frac{(Z_{0.625}) \exp(0.5h_L Z_{0.625}^2) - (Z_{0.375(L)}) \exp(0.5h_L Z_{0.375(L)}^2)}{(Z_{0.7(R)}) \exp(0.5h_R Z_{0.7(R)}^2) - (Z_{0.3}) \exp(0.5h_R Z_{0.3}^2)}\end{aligned}\quad (17)$$

$$\begin{aligned}\alpha_4^R &= \frac{F^{-1}(Z_{0.625}) - F^{-1}(Z_{0.375})}{F^{-1}(Z_{0.7}) - F^{-1}(Z_{0.3})} \\ &= \frac{(Z_{0.625}) \exp(0.5h_R Z_{0.625}^2) - (Z_{0.375(R)}) \exp(0.5h_R Z_{0.375(R)}^2)}{(Z_{0.7(L)}) \exp(0.5h_L Z_{0.7(L)}^2) - (Z_{0.3}) \exp(0.5h_L Z_{0.3}^2)}\end{aligned}\quad (18)$$

$$\begin{aligned}
\alpha_i^5 &= \frac{F^{-1}(Z_{0.5}) - F^{-1}(Z_{0.1(L)})}{F^{-1}(Z_{0.9(R)}) - F^{-1}(Z_{0.5})} \\
&= \frac{(Z_{0.5}) \exp(0.5h_L Z_{0.5}^2) - (Z_{0.1(L)}) \exp(0.5h_L Z_{0.1(L)}^2)}{(Z_{0.9(R)}) \exp(0.5h_R Z_{0.9(R)}^2) - (Z_{0.5}) \exp(0.5h_R Z_{0.5}^2)} \\
&= \frac{-(Z_{0.1(L)}) \exp(0.5h_L Z_{0.1(L)}^2)}{(Z_{0.9(R)}) \exp(0.5h_R Z_{0.9(R)}^2)}
\end{aligned} \tag{19}$$

$$\begin{aligned}
\alpha_5^R &= \frac{F^{-1}(Z_{0.5}) - F^{-1}(Z_{0.1})}{F^{-1}(Z_{0.9}) - F^{-1}(Z_{0.5})} \\
&= \frac{(Z_{0.5}) \exp(0.5h_R Z_{0.5}^2) - (Z_{0.1(R)}) \exp(0.5h_R Z_{0.1(R)}^2)}{(Z_{0.9(L)}) \exp(0.5h_L Z_{0.9(L)}^2) - (Z_{0.5}) \exp(0.5h_L Z_{0.5}^2)} \\
&= \frac{-(Z_{0.1(R)}) \exp(0.5h_R Z_{0.1(R)}^2)}{(Z_{0.9(L)}) \exp(0.5h_L Z_{0.9(L)}^2)}
\end{aligned} \tag{20}$$

$$\begin{aligned}
\alpha_6^l &= \frac{F^{-1}(Z_{0.75}) - F^{-1}(Z_{0.25})}{F^{-1}(Z_{0.9}) - F^{-1}(Z_{0.1})} \\
&= \frac{(Z_{0.75(R)}) \exp\left(\frac{h_R Z_{0.75(R)}^2}{2}\right) - (Z_{0.25(L)}) \exp\left(\frac{h_L Z_{0.25(L)}^2}{2}\right)}{(Z_{0.9(R)}) \exp\left(\frac{h_R Z_{0.9(R)}^2}{2}\right) - (Z_{0.1(L)}) \exp\left(\frac{h_L Z_{0.1(L)}^2}{2}\right)}
\end{aligned} \tag{21}$$

$$\begin{aligned}
\alpha_6^R &= \frac{F^{-1}(Z_{0.75}) - F^{-1}(Z_{0.25})}{F^{-1}(Z_{0.9}) - F^{-1}(Z_{0.1})} \\
&= \frac{(Z_{0.75(L)}) \exp\left(\frac{h_L Z_{0.75(L)}^2}{2}\right) - (Z_{0.25(R)}) \exp\left(\frac{h_R Z_{0.25(R)}^2}{2}\right)}{(Z_{0.9(L)}) \exp\left(\frac{h_L Z_{0.9(L)}^2}{2}\right) - (Z_{0.1(R)}) \exp\left(\frac{h_R Z_{0.1(R)}^2}{2}\right)}.
\end{aligned} \tag{22}$$

In situations where $h_L = h_R$, the HH distribution is symmetric, and consequently the third and fifth cumulants are equal to 1.

4 Discussion

The key theoretical result of this paper is that the fifth-ordered percentile based system parameters can be derived for the Tukey family of HR, HQ, and HH distributions. The existence of closed-form solutions helps the practitioners in simulating these distributions using the MOP estimates, as there is no need for the use of numerical methods.

There is a possibility to develop other methods of transformation using for example the Johnson or Burr system in terms of MOP. It is known from the literature that the MOP approach results in a relatively smaller bias and standard error than the MOM approach. Future studies can examine the comparison of the MOP vs. MOM to further support this conclusion.

References

- Badrinath, S. G., & Chatterjee, S. (1988). On measuring skewness and elongation in common stock return distributions. The case of the market index. *Journal of Business*, 61, 451–472.
- Badrinath, S. G., & Chatterjee, S. (1991). A data-analytic look at skewness and elongation in common stock return distributions. *Journal of Business & Economic Statistics*, 9, 223–233.
- Burr, I. W. (1942). Cumulative frequency functions. *Annals of Mathematical Statistics*, 13(2), 215–232.
- Burr, I. W. (1973). Parameters for a general system of distributions to match a grid of α_3 and α_4 . *Communications in Statistics – Theory and Methods*, 2(1), 1–21.
- Burr, I. W., & Cislak, P. J. (1968). On a general system of distributions: I. Its curve-shape characteristics; II. The sample median. *Journal of the American Statistical Association*, 63, 627–635.
- Fischer M., Horn A., & Klein I. (2003). Tukey-type distributions in the context of financial data. *Diskussionspapiere // Friedrich-Alexander-Universität Erlangen-Nürnberg. Lehrstuhl für Statistik und Ökonometrie*, 52.
- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43, 521–532.
- Goerg, G. (2011). *The Lambert way to Gaussianize skewed, heavy tailed data with the inverse of Tukey's h transformation as a special case*. Cornell University Library.
- Guegan, D., & Hassani, B. (2009). A modified Panjer algorithm for operational risk capital calculations. *Journal of Operational Risk*, 4, 26.
- Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate non-normal distributions. *Computational Statistics and Data Analysis*, 40, 685–711.
- Headrick, T. C. (2010). *Statistical simulation: Power method polynomials and other transformations*. Chapman and Hall/CRC.
- Headrick, T. C., & Pant, M. D. (2012). Characterizing Tukey h and hh-distributions through L-moments and the L-correlation. *ISRN Applied Mathematics*, 2012, 1–20.
- Headrick, T. C., Kowalchuk, R. K., & Sheng, Y. (2008). Parametric probability densities and distribution functions for Tukey g-and-h transformations and their use for fitting data. *Applied Mathematical Sciences*, 2(9), 449–462.
- Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distributions. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Exploring data, tables, trends, and shapes* (pp. 461–511). Wiley.

- Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, *36*, 149–176.
- Karian, Z. A., & Dudewicz, E. J. (2011). *Handbook of fitting statistical distributions with R*. Chapman & Hall.
- Kowalchuk, R. K., & Headrick, T. C. (2010). Simulating multivariate g-and-h distributions. *The British Journal of Mathematical and Statistical Psychology*, *63*, 63–74.
- Kuo, T. C., & Headrick, T. C. (2014). Simulating univariate and multivariate Tukey g-and-h distributions based on the method of percentiles. *ISRN Probability and Statistics*, *2014*, 1–10. <https://doi.org/10.1155/2014/645823>
- Kuo, T. C., & Headrick, T. C. (2017). A characterization of power method transformations through the method of percentiles. Paper presented at the annual meeting of the American Educational Research Association, San Antonio.
- Mills, T. C. (1995). Modelling skewness and kurtosis in the London stock exchange FT-SE index return distributions. *Journal of the Royal Statistical Society: Series D*, *44*, 323–332.
- Morgenthaler, S., & Tukey, J. W. (2000). Fitting quantiles: Doubling, HR, HQ, and HHH distributions. *Journal of Computational and Graphical Statistics*, *9*, 180–195.
- Pant, M., & Headrick, T. C. (2017). A characterization of the Burr Type III and Type XII distributions through the method of percentiles and the Spearman correlation. *Communications in Statistics: Simulation and Computation*, *46*, 1611–1627. <https://doi.org/10.1080/03610918.2015.1048878>
- Pearson, K. (1895). Contributions to the mathematical theory of evolution. III. Regression, heredity, and panmixia. *Proceedings of the Royal Society*, *59*, 69–71.
- Pearson, K. (1901). Mathematical contributions to the theory of evolution. X. Supplement to a memoir on skew variation. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *197*, 443–459.
- Pearson, K. (1916). Mathematical contributions to the theory of evolution. XIX. Second supplement to a memoir on skew variation. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *216*, 429–457.
- Tukey, J. W. (1960). *The practical relationship between the common transformation of percentages of counts and of amounts*. Technical Report 36, Statistical Techniques Research Group, Princeton University.
- Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.
- Yari, G., & Tondpour, Z. (2017). The new Burr distribution and its application. *The Mathematical Scientist*, *11*, 47–54. <https://doi.org/10.1007/s40096-016-0203-z>

Predicting Human Psychometric Properties Using Computational Language Models



Antonio Laverghetta Jr., Animesh Nighojkar, Jamshidbek Mirzakhlov, and John Licato

Abstract Transformer-based language models (LMs) continue to achieve state-of-the-art performance on natural language processing (NLP) benchmarks, including tasks designed to mimic human-inspired “commonsense” competencies. To better understand the degree to which LMs can be said to have certain linguistic reasoning skills, researchers are beginning to adapt the tools and concepts from psychometrics. But to what extent can benefits flow in the other direction? In other words, can LMs be of use in predicting the psychometric properties of test items, when those items are given to human participants? If so, the benefit for psychometric practitioners is enormous, as it can reduce the need for multiple rounds of empirical testing. We gather responses from numerous human participants and LMs (transformer- and non-transformer-based) on a broad diagnostic test of linguistic competencies. We then use the human responses to calculate standard psychometric properties of the items in the diagnostic test, using the human responses and the LM responses separately. We then determine how well these two sets of predictions correlate. We find that transformer-based LMs predict the human psychometric data consistently well across most categories, suggesting that they can be used to gather human-like psychometric data without the need for extensive human trials.

Keywords Classical test theory · Item response theory · Natural language processing

1 Introduction

The current generation of transformer-based language models (TLMs) (Vaswani et al., 2017) continues to surpass expectations, consistently achieving state-of-the-art results on many natural language processing (NLP) tasks. Transformers are a

A. Laverghetta Jr. (✉) · A. Nighojkar · J. Mirzakhlov · J. Licato
Advancing Machine and Human Reasoning (AMHR) Lab, Department of Computer Science and Engineering, University of South Florida, Tampa, FL, USA
e-mail: alaverggett@usf.edu; anighojkar@usf.edu; mirzakhlov@usf.edu; licato@usf.edu

type of artificial neural network that connect text encoders and decoders without using recurrent links, as was the case in previous architectures such as Long Short Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997). Instead, they rely on a computationally efficient self-attention mechanism (Vaswani et al., 2017). Especially surprising is the remarkable performance of these models on benchmark tasks designed to assess “commonsense” reasoning (e.g., Wang et al., 2018, 2019), possibly owing to their ability to encode and retrieve a surprising amount of structural knowledge (Goldberg, 2019; Hu et al., 2020; Cui et al., 2020).

Understanding how TLMs reason is a complex task made more difficult by the fact that the sizes of contemporary TLMs are so large that they are effectively black boxes. As such, researchers are continually searching for new methods to understand the strengths and limitations of TLMs. One promising approach is to draw from the tools of psychometrics, which allows us to measure latent attributes like reasoning skills, even if the mechanisms giving rise to these attributes is not well understood. Although some have called for bridging the gap between psychometrics and artificial intelligence (AI) (Bringsjord, 2011; Bringsjord and Licato, 2012; Hernández-Orallo et al., 2016; Wilcox et al., 2020), the amount of work attempting to do so has been limited. While methods from psychometrics could certainly be useful as a diagnostic tool for AI practitioners, the remarkable performance of TLMs on reasoning tasks suggests that they might also be useful to psychometricians when designing evaluation scales. Most prior work has focused on the benefits psychometrics can bring to AI, however, and has not considered whether tools from AI can also benefit psychometrics, which is the focus of the present paper.

To illustrate how AI might be applied to psychometrics, assume that someone wishes to design a test to assess the degree to which a person possesses mastery of some cognitive skill S . A good place to start is for a panel of experts to design a set of test items \mathcal{I} , such that they believe solving \mathcal{I} requires S , and can therefore be used to measure mastery of S . A common task in psychometrics is to design measurement tools such as \mathcal{I} , and then to apply \mathcal{I} to a large number of human participants. The data obtained from these trials can be used to estimate psychometric properties of the items in \mathcal{I} , such as their reliability, validity, and fairness. But establishing these properties can be prohibitively costly, requiring large numbers of human participants to answer the items in \mathcal{I} and iteratively refine them. This drawback motivates our central research question: **Can TLMs be used to predict psychometric properties of test items?** Psychometrics would benefit greatly if so, as TLMs could be used in place of human participants, reducing the need for extensive human trials.

We present the first exploration into how well TLMs can be used to predict certain psychometric properties of linguistic test items. To do this, we identified a subset of items from the General Language Understanding Evaluation (GLUE) broad coverage diagnostic (Wang et al., 2018), a challenging benchmark of linguistic reasoning skills used to measure the progress of language modeling in the NLP community. We collected human responses on these items to assess simple psychometric properties, designing a novel user validation procedure to do so. We then assess the performance of 240 language models (LMs) on these diagnostic items. Our resulting analysis suggests TLMs show promise in modeling human

psychometric properties in certain sub-categories of linguistic skills, thus providing fruitful directions for future work.

2 Background in Natural Language Processing

As our work draws heavily on models, datasets, and techniques from NLP, we will begin by briefly introducing some important concepts that will be used throughout this work. Note that this is not meant to be an exhaustive introduction to the field; the interested reader is encouraged to refer to the citations throughout this section for more details.

2.1 Language Modeling

In NLP, language models (LMs) are the primary tool used to perform tasks related to natural language understanding (e.g., sentiment analysis, machine translation, and so forth). All the models used throughout this work are examples of LMs. Given a sequence of words, the task of an LM is to predict which word is most likely to come next:

$$P(w_t|w_{1:t-1}) \tag{1}$$

Where w_t is the word to be predicted by the LM at timestep t , and $w_{1:t-1}$ is the prior $t - 1$ words given to the LM to be used to make said prediction (Jurafsky, 2000).

An LM can be constructed using a variety of probabilistic models, however, the one most relevant to this work is the artificial neural network (ANN). ANNs are models from deep learning that consist of three types of units: an input layer, one or more hidden layers, and an output layer. At a high level, they operate by taking in as input a vector representation in the input layer, performing a series of transformations on the input in each hidden layer, and finally mapping the hidden layer to fixed-length representation in the output layer. Figure 1 shows a schematic representation of a simple 2-layer ANN. As the hidden layers within an ANN can perform a variety of non-linear transformations to the input, ANNs are quite expressive in the kinds of representations they can learn (Yarotsky, 2022), which makes them highly effective as models of language. The neural language model was first introduced in Bengio et al. (2003), and works by using ANNs to approximate the probability of each word, given the prior sequence of words.

Since the advent of neural language modeling, more sophisticated neural networks have been employed in NLP, including Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) networks, and transformers (Vaswani et al., 2017). These types of neural networks perform the same basic set of operations as the vanilla ANN, but they differ in how their architectures are designed. LSTMs rely

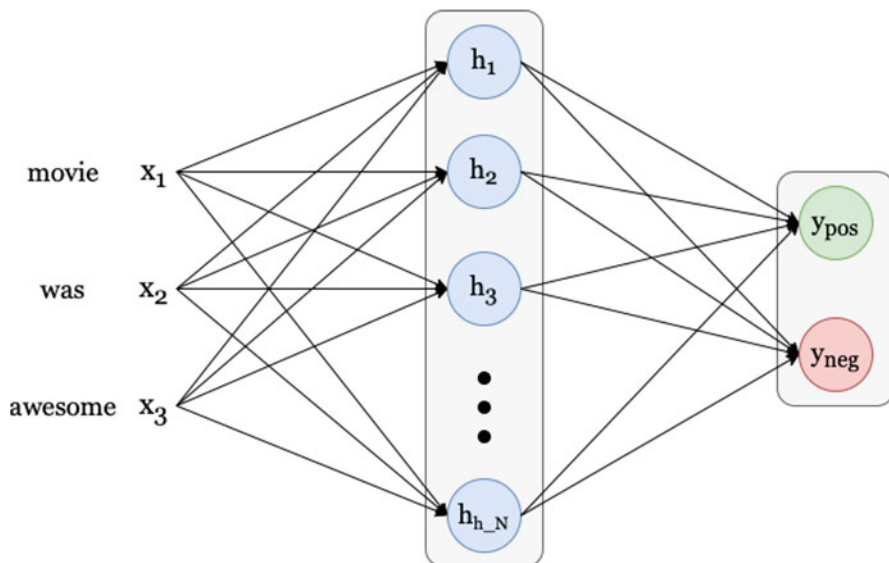


Fig. 1 A simple ANN for the task of sentiment analysis. Words are input to the hidden layers, which learn to map an arbitrary sequence to a fixed output space (positive or negative sentiment). Note that the input layer is typically not counted when listing the total number of layers

on using recurrent (cyclic) links between hidden layers, which allows information from previous hidden layers to affect the representations learned in later layers. Transformers rely on a technique in deep learning called attention, which is meant to mimic the attention employed in human cognitive systems. Attention works by masking out less relevant portions of the input, such that they contribute less information to later layers. For example, given the sentence “The dog sat in the chair.”, attention would learn that “sat” and “chair” contribute more to the meaning of “dog” in this sentence than words like “the.” As discussed earlier, while attention has been employed in previous types of neural networks, transformers are unique in that they use only attention to learn representations of the input, throwing out recurrent layers entirely. Figure 2 shows the general structure of a transformer. A transformer block consists of only an attention operation, followed by a standard hidden layer from a typical ANN. Despite their simplicity, transformers have proven to be highly versatile models, and have surpassed the performance of previous successful architectures on virtually every NLP task.

Training any LM requires that we have access to a large corpus of text, which the LM uses to learn which words frequently occur in which contexts. While there are a variety of approaches to training an LM, by far the most successful of them was pioneered by Devlin et al. (2018) who introduced the BERT (Bidirectional Encoder Representations from Transformers) LM. BERT is a transformer that is trained in two stages, the first being *pre-training* where the model is trained using a self-supervised language modeling objective over a large corpus of text. In the second

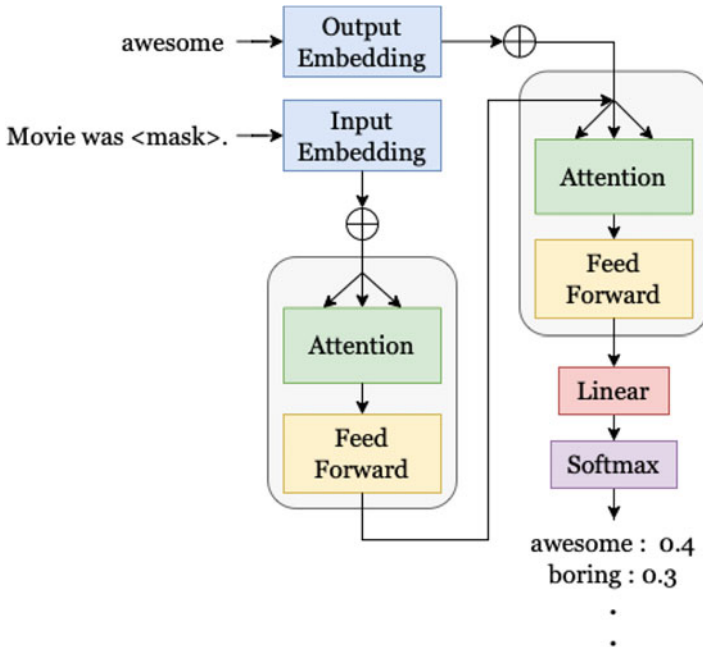


Fig. 2 The architecture of the transformer. The input sequence is given with a mask token <mask> and the correct word is given as output. The model predicts the probability distribution for the mask token and changes its weights after comparing its predictions with the actual next word. During pre-training, this process is repeated many times over a large number of sentences and documents

finetuning stage the model is further trained on a labeled dataset for a particular task, thus allowing the same pre-trained model to be used for many different tasks. One can think of the pre-training stage as giving the model a large amount of domain-general knowledge, whereas the finetuning stage focuses on how to use that knowledge to solve a specific task. Most transformers introduced after BERT use this same training strategy, though the details may differ.

2.2 Natural Language Inference

Natural language inference (NLI) is a common task in NLP for evaluating the reasoning capabilities of LMs. NLI problems consist of two sentences: a premise (p) and hypothesis (h), and solving such a problem involves assessing whether p textually entails h . There are typically three choices: either p does textually entail h (entailment), p entails that h is impossible (contradiction), or h 's truth can not be determined from p alone (neutral). Whether p entails or does not entail h can depend on many factors, such as the syntactic relationships between the sentences,

the information that the sentences convey, or some external knowledge about the world. For example, consider an NLI question with $p =$ “My dog needs to be walked.” and $h =$ “My dogs need to be walked.” We would say that h contradicts p because it was established in p that I have only one dog. As another example, consider $p =$ “The BART line I always take was delayed.” and $h =$ “I’m going to miss my tour of the Statue of Liberty.” We might say that this is a contradiction because the BART operates in San Francisco and not New York City. However, we might also say that p is neutral with respect to h (perhaps I need to ride the BART to the airport, where I will then fly to New York City). Regardless, this demonstrates how the NLI task can also incorporate external information not explicitly stated in either sentence.

The NLI task was formalized in the PASCAL recognizing textual entailment tasks (Dagan et al., 2006), which were a series of workshops designed to spur the development of NLP systems for inferential reasoning. The NLI datasets developed for these tasks were quite small, having only a few thousand items in total, which made it very difficult to train deep neural networks on them. The Stanford natural language inference (SNLI) (Bowman et al., 2015) corpus was the first large-scale dataset of NLI questions, having around 570,000 items in total, which made it practical to train LMs for NLI. Since the release of SNLI, other large-scale NLI datasets have been curated, including MultiNLI (MNLI) (Williams et al., 2018) and Adversarial NLI (ANLI) (Nie et al., 2020), each of which curates NLI questions of varying levels of difficulty and covers different domains of text (fictional stories, news, telephone conversations, etc.). This has made the NLI task quite general in the kinds of reasoning it can test for, while also being straightforward to administer to both humans and LMs, which makes the task ideal for the present study.

2.3 Benchmarks of Commonsense Reasoning

A common task in NLP is the development of tasks and datasets meant to assess the language understanding and reasoning capabilities of new models. Such tasks are typically narrowly scoped, focusing on how well the model performs on one specific task. More recently, there has been a trend to developing more comprehensive assessments of LM performance, meant to mimic the diverse skill sets a model would need to master when operating in the real world. The General Language Understanding Evaluation (GLUE), as well as its more recent extension SuperGLUE (Wang et al., 2018, 2019), are such benchmarks and are meant to assess a broad set of linguistic reasoning competencies. GLUE was curated by combining previous datasets into a single benchmark task, covering a diverse set of underlying skills, including NLI, question answering, paraphrase detection, and others. As there has been rapid progress in NLP in recent years, the authors of GLUE found that the benchmark quickly lost the ability to discriminate between high and low-performance LMs on the tasks it covered. SuperGLUE (Wang et al., 2019) was then curated to address this, using a newer suite of more challenging tasks.

Most relevant to this work is the GLUE task known as the broad coverage diagnostic, which is a set of items formatted as NLI problems. The diagnostic covers four main categories of linguistic competencies: *lexical semantics*, *predicate-argument structure*, *logic*, and *knowledge and common sense*. These categories are further divided into multiple sub-categories, each of which covers a specific and interesting phenomenon in language. The broad coverage diagnostic was manually curated by linguistics and NLP experts and is meant to assess broad psycholinguistic competencies of LMs across multiple categories. For instance, the *propositional structure* category contains questions that exploit propositional logic operators, e.g., p = “The cat sat on the mat.” and h = “The cat did not sit on the mat.” The diagnostic thus aims to be a comprehensive test of linguistic reasoning skills, making it suitable for our present study. As discussed in Sect. 3, we use only the following seven sub-categories from the diagnostic for our experiments:

1. *morphological negation*: Covers questions that require reasoning over negation in either its logical or psycholinguistic form.
2. *prepositional phrases*: Tests for the ability to handle ambiguity introduced by the insertion or removal of prepositions (e.g., p = “Cape sparrows eat seeds, along with soft plant parts and insects.” and h = “Soft plant parts and insects eat seeds.”).
3. *lexical entailment*: Covers hypernymy, hyponymy, and other types of monotonic relationships at the word level (e.g., a dog is an animal, but is not a cat).
4. *quantifiers*: Tests for the ability to reason over the universal and existential logical operators.
5. *propositional structure*: Tests for the ability to reason over the core suite of logical operators, including conjunction, disjunction, and conditionals.
6. *richer logical structure*: Covers higher-level forms of logic, especially those dealing with temporal or numeric reasoning.
7. *world knowledge*: Tests for knowledge of specific factual information about the world.

3 Gathering Language Model Data

We begin by gathering results on the broad coverage diagnostic from a suite of LMs. We first selected a subset of the diagnostic items that were a member of only one sub-category, to better isolate factors. From this subset, we had 811 diagnostic questions encompassing 20 sub-categories. Each sub-category had at least 15 questions, and we selected the seven sub-categories enumerated in Sect. 2.3 to use in our experiments. We selected these 7 sub-categories based on how much the average performance of the LMs improved after pre-training and finetuning. A substantial performance improvement indicated the category was solvable by the models, and would therefore provide a meaningful comparison to the human data.

We gathered responses to the diagnostic from a wide array of TLMs, including BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), T5 (Raffel et al., 2020), ALBERT (Lan et al., 2020), XLNet (Yang et al., 2019), ELECTRA (Clark et al., 2020), Longformer (Beltagy et al., 2020), SpanBERT (Joshi et al., 2020), DeBERTa (He et al., 2020), and ConvBERT (Jiang et al., 2020). Each of these models differs from the others along one or more factors, including underlying architecture, pre-training objective and data, or the general category the model belongs to. We experimented with multiple different “snapshots” of each TLM. We obtained these snapshots from HuggingFace (Wolf et al., 2020). For each model we used a smaller version, designated with the *small* or *base* suffix, and a larger version, designated with the *base* or *large* suffix. The smaller versions of each TLM contained fewer transformer blocks, and thus fewer trainable parameters, making them less expressive models of language. We used LSTM-based LMs (Hochreiter and Schmidhuber, 1997) as a baseline, which, unlike TLMs, primarily rely on recurrent links, as opposed to attention.

We used the SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), and ANLI (Nie et al., 2020) datasets to finetune our models for the NLI task. To increase the variance in our results as much as possible, we finetuned all models on various combinations of these datasets: (1) SNLI alone, (2) MNLI alone, (3) SNLI + MNLI, and (4) SNLI + MNLI + ANLI. Recall that all TLMs are trained in two stages: pre-training and then finetuning. As the performance of our models on the diagnostic will be affected by both, we systematically alter whether a model is pre-trained or finetuned to further increase variance, using the following combinations:

- **Zero shot:** The model is initialized with random weights in the hidden layers and is evaluated on the diagnostic without any training. This is meant to test whether there is any property of the architecture itself which is useful for solving the diagnostic.
- **Pre-train, no finetune:** The model is pre-trained but not finetuned. In this case, the language model is still fully trained, but it has not been specifically optimized for NLI.
- **No pre-train, finetune:** The model weights are initialized randomly, but we finetune the model before evaluating it. The model is trained for NLI, but the total amount of language it has been exposed to is much smaller without pre-training.
- **Pre-train and finetune:** The model is fully trained before evaluation.

For BERT, we experimented with both Devlin et al. (2018)’s pre-trained models, and a BERT model we trained from scratch. Our BERT model had an identical architecture to *bert-base* and was pre-trained on Google’s One Billion Words corpus (Chelba et al., 2014), which is a dataset of documents from various sources created by Google for pre-training LMs.

In summary, this process allowed us to vary the underlying architecture, the size of each architecture, and the amount of data the model was trained on. This allowed us to treat each trained model as effectively being a different “individual” (and we will refer to them as such), which might have a radically different cognitive profile

from its counterparts. For example, a *roberta-base* model that was pre-trained and finetuned on all three NLI datasets would likely be much more proficient on our diagnostic than a *roberta-large* model trained on no NLI data at all.

4 Human Studies

As our purpose in gathering this LM data was to evaluate it against human performance, we additionally ran a human study. To do this, we recruited workers on Amazon Mechanical Turk (mTurk) to complete our subset of GLUE diagnostic questions. While mTurk makes conducting large-scale human studies convenient, there are also well-documented problems with participants not completing tasks in good faith (Berinsky et al., 2014). There are multiple techniques for filtering out bad-faith participants, such as the use of *attention check* questions, sometimes called “instructional manipulation checks” (Hauser and Schwarz, 2015), which are designed so that a good-faith participant would be unlikely to get them incorrect. But this alone would not suffice for our purposes here, as we want a certain amount of low-scoring participants on some sub-categories, so that the population variances on sub-category items would better reflect their actual variances.

We first obtained attention checks from the ChaosNLI dataset (Nie et al., 2020), which gathered over 450,000 human annotations on questions from SNLI and MNLI. Since each question in ChaosNLI was annotated by 100 different workers, if the inter-annotator agreement for a given question is extremely high, we conclude that question is likely easy to solve for good-faith participants. We gathered 36 questions from ChaosNLI where the agreement for the correct label was at least 90%. These were enough questions to ensure that each phase of our trials used a unique set of attention check questions. The human studies were split up into five phases, and workers who did sufficiently well in a given phase were given a qualification to continue to the next phase:

1. **On-boarding:** A qualifying HIT (human intelligence task) open to any worker located in the United States, who had completed at least 50 HITs with an approval rating of at least 90%. The HIT consisted of five attention check questions, given to each worker in the same order. We gathered up to 200 responses and paid workers \$0.50.
2. **Phase 1:** Included questions from *morphological negation*, and three attention checks. We gathered up to 45 responses and paid workers \$3.60.
3. **Phase 2:** Included questions from *lexical entailment* and *prepositional phrases*, as well as six attention checks. We gathered up to 36 responses and paid workers \$7.20.
4. **Phase 3:** Included questions from *quantifiers* and *propositional structure*, as well as six attention checks. We gathered up to 27 responses and paid workers \$7.20.

5. **Phase 4:** Included questions from *richer logical structure* and *world knowledge*, as well as six attention checks. We gathered responses from all accepted workers from Phase 3 and paid workers \$7.20.

Our payment structure was designed to incentivize workers to put forth their best effort when completing the task. Workers were informed that successfully completing each task would award them the opportunity to earn additional payment on each subsequent phase. However, if on a given phase a worker failed our authentication protocol (described below), we rejected their work and did not pay them. Workers were informed before starting every study that we would evaluate the quality of their work, and that it might be rejected if we found evidence that they did not put forth an honest effort.

In each phase, questions were randomly ordered, except for attention checks which were spread evenly throughout the survey. We used Qualtrics¹ to create the surveys for each HIT and collect the responses. Participants were first presented with instructions for the task and some examples, which were based on the instructions originally given to annotators of the MNLI dataset. The questions from each category were a randomly chosen subset of 15 questions tested on the LMs for that category, balanced for each label. For each question, workers also had to provide a short justification statement on why they believed their answer was correct, which was used to help filter out bad faith participants. To validate the responses to our surveys, we developed the following authentication procedure:

1. Look for duplicate IPs or worker IDs, indicating that the worker took the HIT more than once. If there are any, keep only the first submission.
2. If the worker's overall score was less than 40%, reject the HIT. If their overall score was greater than 60%, accept the HIT. For workers who scored between 40 and 60%, reject the HIT if they got less than 75% of the attention checks correct.
3. Finally, examine the justifications of all workers not previously rejected. Here we were looking for simple, but clear, reasons for why workers chose their answer. We included this step because we found in a pilot study that workers sometimes provided nonsensical justifications for their answers even when they did well on the survey, making it unclear whether they were truly paying attention. We checked that the justifications appeared relevant to the question, that they did not paste part of the question for their justification, that they did not use the same justification for every question, and that they did not use short nonsensical phrases for their justification (e.g., some simply wrote "good" or "nice" as their justification). This allowed us to keep some low-scoring participants who had put genuine effort into the task.

Manual inspection of the resulting responses suggested that workers whose responses were accepted consistently gave higher quality responses than those who did not. These workers gave more detailed justifications that clearly articulated their

¹ <https://www.qualtrics.com>.

thought process, often citing specific details from the question. On the other hand, workers who failed to give good justifications generally scored at or below random chance, which further indicated that they were not actually paying attention. We, therefore, believe the use of justifications helped us gather higher-quality responses.

Using this procedure, and those described in Sect. 3, we gathered results from 27 human participants and 240 neural LMs (183 transformer-based and 57 LSTM-based). In addition to the LSTMs, we also include a true random baseline which simply guesses randomly on every question. In the following experiments, we use the human performance on each category as the basis for analyzing the performance of the artificial populations, specifically using methods from classical test theory (both simple problem difficulty and inter-item correlation) and Rasch models (Rasch, 1993) from item response theory. Our goal is to determine how well item properties measured using artificial models correlate with those measured using the humans responses, using both pearson and spearman correlation coefficients. We shall refer to the transformer population as T , the LSTM population as L , the random population as R , and the human population as H . We used the ltm R package to fit all Rasch models (Rizopoulos, 2006).

5 Experimental Results

5.1 Classical Test Theory

We began by examining how well TLMs could predict simple problem difficulty in the human data. For each item i in a given sub-category, we calculated the percentage of human participants who got that item correct (D_H^i), and then the corresponding percentage for the TLMs (D_T^i), LSTM-based LMs (D_L^i), and the random baseline (D_R^i). We then calculated the Spearman correlation between D_H^i and each of the other populations. Results are shown in Table 1. In almost all cases,

Table 1 Given D_H , Spearman correlation and p -values were calculated with transformer-based (D_T), LSTM-based (D_L), and random (D_R) estimates of problem difficulty. Note that we have bolded cells whose correlations (absolute values) were highest, but their p -values were not always significant. Columns marked with * are significant at $p < 0.05$, ** at $p < 0.01$, and *** at $p < 0.001$

Category	D_T	D_L	D_R
Morphological negation	-0.28	0.27	-0.14
Prepositional phrases	*** 0.86	0.47	0.42
Lexical entailment	* 0.62	0.17	-0.22
Quantifiers	* 0.57	-0.22	0.41
Propositional structure	*** 0.93	0.27	0.37
Richer logical structure	0.28	-0.03	-0.37
World knowledge	*** 0.79	0.46	-0.25

Table 2 Pearson correlation and p -values for how well items clustered using human responses match the clusters which used transformer-based (C_T), LSTM-based (C_L), and random (C_R) items. Columns marked with * are significant at $p < 0.05$, ** at $p < 0.01$, and *** at $p < 0.001$

Category	C_T	C_L	C_R
Morphological negation	0.18	*** 0.40	-0.14
Prepositional phrases	** 0.31	-0.15	-0.01
Lexical entailment	** 0.31	-0.03	-0.16
Quantifiers	* 0.24	-0.01	0.06
Propositional structure	*** 0.51	0.03	0.04
Richer logical structure	*** 0.46	-0.07	0.04
World knowledge	** 0.28	0.00	-0.09

TLMs achieve a much stronger correlation with the human data than either baseline. The main exceptions are *morphological negation* and *richer logical structure*, both of which fail to produce strong, statistically significant correlations. As we will see, this pattern will repeat in other measurements as well.

IIC-Based Clustering An important idea in psychometrics is that items that rely on the same skills should have similar chances of being answered correctly by a given participant (Rust and Golombok, 2014). Whether items rely on similar skills can be tested using the inter-item correlation (IIC) between two items, where high IIC suggests that the items rely on similar underlying reasoning skills. Thus, it can be assumed that if items cluster together when using IIC as a distance metric, they rely on similar underlying cognitive skills. To explore this, given a correlation measure c ranging from -1 to 1 , we converted it into a distance metric by taking $1 - c$. We used this metric to cluster the diagnostic questions. For each sub-category, we performed clustering using human, transformer, LSTM, and random data separately (H , T , L , and R respectively).

After clustering, for each pair of items (i, j) we define $C_{i,j}^D$ as 1 if i and j are in the same cluster as determined by dataset $D \in \{H, T, L, R\}$, and 0 otherwise. Finally, to determine how well clusters from the LM responses match the human responses, we calculated Pearson correlation between C_H and each of C_T , C_L , and C_R . Results are shown in Table 2. Similar to Table 1, we see statistically significant correlations from TLMs in every sub-category, except again for *morphological negation*.

5.2 Item Response Theory

Since TLMs correlated well with humans using the classical techniques we tested, we wished to examine whether this would still hold using methods from item response theory (IRT). To do this, we used the diagnostic results from each population to fit Rasch models (Rasch, 1993). This gave us separate difficulty parameter estimates b_i for each item i , for each population. To determine how well the difficulty parameters matched between populations, we calculated the

Table 3 Pearson correlation and p -values for transformer-based (D_T), LSTM-based (D_L), and random (D_R) estimates of problem difficulty computed using Rasch models. Columns marked with * are significant at $p < 0.05$, ** at $p < 0.01$, and *** at $p < 0.001$

Category	D_T	D_L	D_R
Morphological negation	0.08	0.29	0.19
Prepositional phrases	0.48	**0.69	-0.25
Lexical entailment	***0.88	-0.06	0.14
Quantifiers	*0.61	0.03	0.12
Propositional structure	*0.61	0.05	-0.25
Richer logical structure	0.16	-0.05	-0.31
World knowledge	*0.52	*0.59	-0.1

Pearson correlation between the b_i using our human response data (H), and the b_i obtained using the other populations (T , L , R). Results are shown in Table 3. As before, TLMs consistently get a stronger correlation than either baseline on most sub-categories, except for *morphological negation* and *richer logical structure*. Interestingly, LSTM-based LMs achieved stronger correlations than TLMs on certain sub-categories: *world knowledge* and *prepositional phrases*. The only other experiment where LSTM-based LMs achieved stronger correlation was reported in Table 2, where they achieved superior correlation on *morphological negation*.

6 Related Work

What reason do we have to suspect that TLMs can predict the psychometric properties of test items? Although TLMs were not primarily designed to compute in a human-like way, there are some reasons to suspect that they may have the ability to effectively model at least some aspects of human linguistic reasoning: They consistently demonstrate superior performance (at least compared to other LMs) on human-inspired linguistic benchmarks (Wang et al., 2018, 2019), and they are typically pre-trained using a lengthy process designed to embed deep semantic knowledge, resulting in efficient encoding of semantic relationships (Zhou et al., 2020; Cui et al., 2020). Common optimization tasks for pre-training transformers, such as the masked LM task (Devlin et al., 2018) are quite similar to the word prediction tasks that are known to predict children’s performance on other linguistic skills (Gambi et al., 2020). Finally, TLMs tend to outperform other LMs in recent work modeling human reading times, eye-tracking data, and other psychological and psycholinguistic phenomena (Schrimpf et al., 2020b,a; Hao et al., 2020; Merx and Frank, 2021; Laverghetta Jr. et al., 2021).

Despite the potential benefits psychometrics could bring to AI, work explicitly bridging these fields has been limited. Ahmad et al. (2020) created a deep learning architecture for extracting psychometric dimensions related to healthcare, specifically numeracy, literacy, trust, anxiety, and drug experiences. Their architecture did not use transformers and relied instead on a sophisticated combination of convolutional and recurrent layers in order to extract representations of emotions,

demographics, and syntactic patterns, among others. Eisape et al. (2020) examined the correlation between human and LM next-word predictions and proposed a procedure for achieving more human-like cloze probabilities. In NLP, methods from IRT have been particularly popular. Lalor et al. (2018) used IRT models to study the impact of item difficulty on the performance of deep models on several NLP tasks. In a follow-up study, Lalor and Yu (2020) used IRT models to estimate the competence of LSTM (Hochreiter and Schmidhuber, 1997) and BERT models during training. Sedoc and Ungar (2020) used IRT to efficiently assess chat-bots. Martínez-Plumed et al. (2019) used IRT to analyze the performance of machine learning classifiers in a supervised learning task. IRT has also been used to evaluate machine translation systems (Otani et al., 2016) and speech synthesizers (Oliveira et al., 2020). Recent work has also used IRT models to evaluate progress on benchmark NLP tasks (Vania et al., 2021; Rodriguez et al., 2021). We contribute to this literature by providing what is, to our knowledge, the first comprehensive assessment of the relationships between human and LM psychometric properties on a broad test of linguistic reasoning.

7 Conclusion

Overall, we find that TLMs perform consistently better than either of our baselines in modeling human psychometric properties. However, this improvement is also not uniform across all categories. In fact, we have found some regularities in this regard. In particular, TLMs failed to achieve a strong correlation on *morphological negation* in all cases. This might be explained by two facts: there is little relative variance in the human responses in this sub-category, and the average accuracy of human participants was above 90%, as opposed to LM accuracy of 55%.

The strong correlation TLMs consistently achieved suggests they can produce similar responses to human participants on diagnostic items. This has many implications for psychometrics, notably the possibility of using them as a sort of simulated test taker for building evaluation scales. If this were successful, it would greatly reduce the burden of multiple rounds of empirical testing.

Of course, this study also has some important limitations. The number of human participants in our study was somewhat small compared to typical psychometrics studies (which often contain hundreds or thousands of participants), making it difficult to draw stronger conclusions. As stated earlier, practical limitations on population size is a common problem in psychometrics research, one which our present work hopes to alleviate somewhat. Future work will need to repeat our experiments with much larger population sizes, and also take measures to ensure sufficient diversity in the study population (e.g., age, income, education level, English fluency, etc.). Furthermore, although we reported in detail on certain psychometrics measures where our method demonstrated promising results for TLMs, it is worth reporting that certain other measures we examined did not appear to align well. For example, item-total correlations using human data did not appear

to correlate with any LM data better than with the random baseline. Likewise, our LMs failed to predict average inter-item correlations between either random subsets of items or our diagnostic sub-categories. More work is needed to better understand why.

While this study has given us some insights into which fundamental reasoning skills TLMs can model well, it does not tell us anything about the *order* in which these skills are acquired, and especially whether this order is at all human-like. For example, in our experiments, we found that TLMs consistently achieved a strong correlation on items requiring mastery of logical operators and lexical entailment (e.g., $p =$ “The dog is on the mat and the cat is in the hat” and $h =$ “The dog is on the mat”). However, if we found that TLMs develop the ability to solve problems with conjunct-containing sentences before those with simpler sentences (e.g., $p =$ “The dog is on the mat” and $h =$ “The dog is not on the mat”) this would clearly not reflect the order of skill acquisition we would expect to see in humans. Other methods from psychometrics, especially cognitive diagnostic models (Rupp and Templin, 2008) might give us a more nuanced understanding of how effective TLMs are as a model of human learning and development.

Finally, while our experiments have given us some insights into the validity and reliability of the diagnostic items, it is unclear whether our approach can allow us to measure their fairness. It is not known whether the test items we examine here are consistent across different groups of differing socio-economic statuses, and we did not control for this in our recruitment. Being able to probe this property of items would have interesting downstream applications. For instance, it might indicate whether a diagnostic gives an unfair advantage to certain types of classifiers, and thus might discriminate against certain groups.

We believe our work offers a clear path forward for bridging psychometrics and AI. The use of psychometric measures gives us a more nuanced understanding of the latent abilities of LMs than single-valued measures like accuracy or F_1 can provide. Furthermore, the increasingly powerful ability of TLMs to model human “commonsense” reasoning and knowledge suggests new ways to predict psychometric properties of test items, reducing the need for costly human empirical data.

Acknowledgments This material is based upon work supported by the Air Force Office of Scientific Research under award numbers FA9550-17-1-0191 and FA9550-18-1-0052. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Air Force.

References

- Ahmad F., Abbasi, A., Li, J., Dobolyi, D. G., Netemeyer, R. G., Clifford, G. D., & Chen, H. (2020). A deep learning architecture for psychometric natural language processing. *ACM Transactions on Information Systems (TOIS)*, 38(1), 1–29.
- Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer. *Preprint arXiv:200405150*

- Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *The Journal of Machine Learning Research*, 3, 1137–1155.
- Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the Shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, 58(3), 739–753.
- Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics.
- Bringsjord, S. (2011) Psychometric artificial intelligence. *Journal of Experimental & Theoretical Artificial Intelligence*, 23(3), 271–277.
- Bringsjord, S., & Licato, J. (2012). Psychometric artificial general intelligence: the piaget-macguyver room. In *Theoretical foundations of artificial general intelligence* (pp. 25–48). Springer.
- Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2014). One billion word benchmark for measuring progress in statistical language modeling. In *Fifteenth Annual Conference of the International Speech Communication Association*.
- Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR 2020: Eighth International Conference on Learning Representations*.
- Cui, L., Cheng, S., Wu, Y., & Zhang, Y. (2020). Does bert solve commonsense task via commonsense knowledge? *Preprint arXiv:200803945*.
- Dagan, I., Glickman, O., & Magnini, B. (2006). The pascal recognising textual entailment challenge. In J. Quiñero-Candela, I. Dagan, & B. Magnini, F. d'Alché Buc (Eds.), *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment* (pp. 177–190). Berlin: Springer.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. N. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1 (Long and Short Papers), pp. 4171–4186).
- Eisape, T., Zaslavsky, N., & Levy, R. (2020). Cloze distillation improves psychometric predictive power. In *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 609–619).
- Gambi, C., Jindal, P., Sharpe, S., Pickering, M. J., & Rabagliati, H. (2020). The relation between preschoolers' vocabulary development and their ability to predict and recognize words. *Child Development*. <https://doi.org/10.1111/cdev.13465>. <https://srcd.onlinelibrary.wiley.com/doi/abs/10.1111/cdev.13465>
- Goldberg, Y. (2019). Assessing bert's syntactic abilities. *CoRR* abs/1901.05287, <http://arxiv.org/abs/1901.05287>, 1901.05287
- Hao, Y., Mendelsohn, S., Sterneck, R., Martinez, R., & Frank, R. (2020). Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 75–86).
- Hauser, D. J., & Schwarz, N. (2015). It's a Trap! Instructional manipulation checks prompt systematic thinking on “Tricky” tasks. *SAGE Open*, 5(2).
- He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with disentangled attention. 2006.03654.
- Hernández-Orallo, J., Martínez-Plumed, F., Schmid, U., Siebers, M., & Dowe, D. L. (2016). Computer models solving intelligence test problems: Progress and implications. *Artificial Intelligence*, 230, 74–107. <https://doi.org/10.1016/j.artint.2015.09.011>. <http://www.sciencedirect.com/science/article/pii/S0004370215001538>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

- Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. (2020). A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1725–1744) Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.158>. <https://www.aclweb.org/anthology/2020.acl-main.158>
- Jiang, Z. H., Yu, W., Zhou, D., Chen, Y., Feng, J., & Yan, S. (2020). ConvBERT: Improving BERT with span-based dynamic convolution. *Advances in Neural Information Processing Systems*, 33, 12837–12848.
- Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8, 64–77.
- Jurafsky, D. (2000). *Speech & language processing*. Pearson Education India.
- Lalor, J. P., Wu, H., Munkhdalai, T., & Yu, H. (2018). Understanding deep learning performance through an examination of test set difficulty: A psychometric case study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, NIH Public Access* (Vol. 2018, p. 4711).
- Lalor, J. P., & Yu, H. (2020). Dynamic data selection for curriculum learning via ability estimation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, NIH Public Access* (Vol. 2020, p. 545).
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. In *ICLR 2020: Eighth International Conference on Learning Representations*.
- Laverghetta Jr, A., Nigohjkar, A., Mirzakhlov, J., & Licato, J. (2021). Can transformer language models predict psychometric properties? In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics* (pp. 12–25) Online. Association for Computational Linguistics.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *Preprint arXiv:1907.11692*.
- Martínez-Plumed, F., Prudêncio, R. B., Martínez-Usó, A., & Hernández-Orallo, J. (2019). Item response theory in AI: Analysing machine learning classifiers at the instance level. *Artificial Intelligence*, 271, 18–42.
- Merkx, D., & Frank, S. L. (2021). Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*(pp. 12–22) Online. Association for Computational Linguistics.
- Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2020). Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Nie, Y., Zhou, X., & Bansal, M. (2020). What can we learn from collective human opinions on natural language inference data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 9131–9143).
- Oliveira, C. S., Tenório, C. C., & Prudêncio, R. (2020). Item response theory to estimate the latent ability of speech synthesizers. In *24th European Conference on Artificial Intelligence - ECAI 2020*.
- Otani, N., Nakazawa, T., Kawahara, D., & Kurohashi, S. (2016). Irt-based aggregation model of crowdsourced pairwise comparison for evaluating machine translations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 511–520).
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67.
- Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. ERIC.
- Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of statistical software*, 17(5), 1–25.

- Rodriguez, P., Barrow, J., Hoyle, A. M., Lalor, J. P., Jia, R., & Boyd-Graber, J. (2021). Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Vol. 1: Long Papers, pp. 4486–4503) Online. Association for Computational Linguistics.
- Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, 6(4), 219–262.
- Rust, J., & Golombok, S. (2014). *Modern psychometrics: The science of psychological assessment*. Routledge.
- Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J., & Fedorenko, E. (2020a). Artificial neural networks accurately predict language processing in the brain. bioRxiv. <https://doi.org/10.1101/2020.06.26.174482>. <https://www.biorxiv.org/content/early/2020/06/27/2020.06.26.174482>. <https://www.biorxiv.org/content/early/2020/06/27/2020.06.26.174482.full.pdf>
- Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J., & Fedorenko, E. (2020b). The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. bioRxiv <https://doi.org/10.1101/2020.06.26.174482>. <https://www.biorxiv.org/content/early/2020/10/09/2020.06.26.174482>. <https://www.biorxiv.org/content/early/2020/10/09/2020.06.26.174482.full.pdf>
- Sedoc, J., & Ungar, L. (2020). Item response theory for efficient human evaluation of chatbots. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems* (pp. 21–33).
- Vania, C., Htut, P. M., Huang, W., Mungra, D., Pang, R. Y., Phang, J., Liu, H., Cho, K., & Bowman, S. R. (2021). Comparing test sets with item response theory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Vol. 1: Long Papers, pp. 1141–1158) Online. Association for Computational Linguistics.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17* (pp. 6000–6010). Red Hook, NY, USA: Curran Associates Inc.
- Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of NeurIPS*.
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 353–355). Brussels, Belgium: Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-5446>. <https://www.aclweb.org/anthology/W18-5446>
- Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. 2006.01912.
- Williams, A., Nangia, N., & Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1 (Long Papers), pp. 1112–1122). Association for Computational Linguistics.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., ... & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45) Online. Association for Computational Linguistics. <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.
- Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems* (Vol. 32, pp 5753–5763).

- Yarotsky, D. (2022). Universal Approximations of Invariant Maps by Neural Networks. *Constructive Approximation*, 55, 407–474. <https://doi.org/10.1007/s00365-021-09546-1>
- Zhou, X., Zhang, Y., Cui, L., & Huang, D. (2020). Evaluating commonsense in pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, pp. 9733–9740).

Predicting Item Characteristic Curve (ICC) Using a Softmax Classifier



Dmitry I. Belov

Abstract The objective of item difficulty modeling (IDM) is to predict the statistical parameters of an item (e.g., difficulty) based on features extracted directly from the item (e.g., number of words). This paper utilizes neural networks (NNs) to predict a discrete item characteristic curve (ICC). The presented approach exploits one-to-one mapping from monotonically non-decreasing discrete ICCs to probability mass functions (PMFs). An NN was trained using soft labels for each item (by mapping ICCs to PMFs), with a softmax output layer representing PMF and the Kullback-Leibler divergence representing a loss function. Results of a cross-validation of the NN on 1742 retired logical reasoning items from the Law School Admission Test are presented and discussed.

Keywords Item difficulty modeling · Item response modeling · Item characteristic curve · Neural networks · Machine learning · Natural language processing · Semantic similarity

1 Introduction

The primary task of item difficulty modeling (or, perhaps more appropriately, item response modeling) is to predict the statistical properties of an item, such as difficulty, based on features extracted directly from the item. An example of such a feature might be the number of words in the item. Item difficulty modeling (IDM) adopts various techniques from data mining, machine learning, and natural language processing. For a review of IDM and its applications see, for example, Sheehan and Mislevy (1990), Huff (2006), or Ferrara et al. (2021).

Due to the recent massive migration of high-stakes testing programs from in-person testing to online testing, the following two issues became much harder to

D. I. Belov (✉)
Law School Admission Council, Newtown, PA, USA
e-mail: dbelov@lsac.org

address without IDM. First, online test proctoring cannot protect against existing technology used to steal test content. In the world of online testing, item preknowledge may happen due to (a) using the same test section over different time slots within the same administration due to a limited number of live proctors; and (b) pretesting new items. Second, a larger number of test sections is needed to tackle the problem set forth in (a). However, developing new items without controlling their statistical parameters may unbalance the pool and limit the assembly of more test sections, thus decreasing item pool usability.

A recent meta-analysis by Ferrara et al. (2021), compiled from over 100 IDM-related studies, demonstrated the following. Only about 10% of the studies reported the coefficient of determination D over 0.5. Most of the research dealt with reading comprehension (RC) items (commonly associated with a long text). The most popular item question concerned the main idea of the passage. All methods predicted only item difficulty. The majority of prediction models utilized linear regression or a regression tree. Features defined by item writers (i.e., nonautomatic features) were often the best predictors of item difficulty.

The current paper goes beyond the typical IDM research described above and instead applies neural networks (NNs) to predict item characteristic curves (ICCs) for logical reasoning (LR) items (which are associated with a shorter amount of text compared to RC items) from the Law School Admission Test (LSAT). For a given item, its ICC maps the examinee's latent trait (ability) to the probability that the item will be answered correctly (Lord, 1980). The ICC is bounded between 0 and 1, is monotonically non-decreasing, and is commonly assumed to take the shape of a logistic function.

This paper considers discrete ICCs defined on the set of ability levels $\{-3, -2, -1, 0, 1, 2, 3\}$ (a coarse grid was chosen just for the sake of illustration; a finer grid is easily supported). There are at least three advantages of predicting discrete ICCs. First, one can avoid the noise produced by the item parameter estimation procedure, while fitting empirical ICCs with an IRT model, by dealing directly with empirical ICCs (see Fig. 1). Second, ICCs provide unification when the item pool has a mixture of models (e.g., part of the item pool modeled by the two-parameter logistic model [2PLM] and the other part by the 3PLM; Lord, 1980): all parts can be represented by ICCs computed using corresponding models. Third, once discrete ICCs are predicted, it is easy to simulate responses from any targeted population of examinees and then calibrate IRT models (1PLM, 2PLM, or 3PLM), thus providing continuous ICCs.

This paper is organized as follows. First, the construction, training, and validation of a neural network (NN) to predict ICCs are described. Second, the data from retired LR items from the LSAT and features extracted from each item are depicted. Third, the results of applying the developed NN to the data are presented. Finally, the results are critically reviewed; followed by a discussion about further research, design changes, and practical applications.

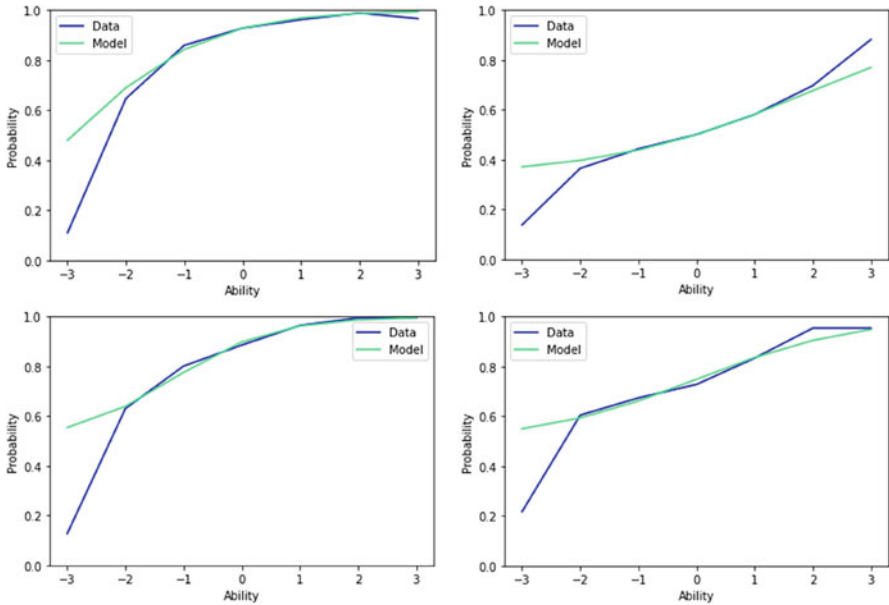
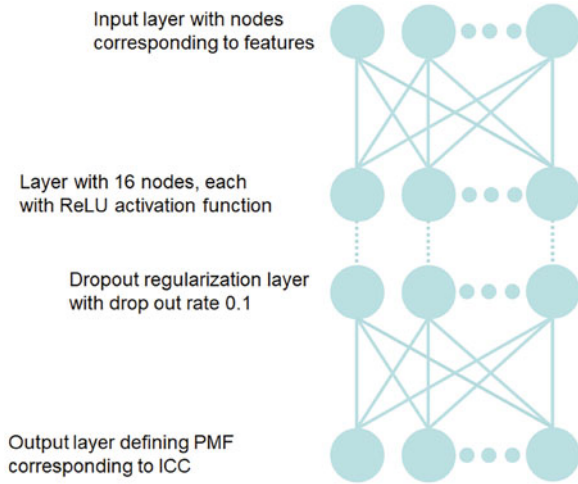


Fig. 1 When empirical ICCs are fitted by item response theory (IRT) models, there is always a possibility of misfit. Here are four real cases showing a large misfit between empirical ICCs (Data) and ICCs produced by a fitted three-parameter logistic (3PL) model (Model)

2 Method

A neural network (NN) can be considered a vector function with a vector argument. In this study, the NN maps the vector of features extracted from an item to its ICC. Parameters of this function can be estimated using a “training” sample, where for each argument there is a predetermined output of the function called a label, by minimizing a loss function. In this process, called *supervised learning*, it is crucial that the training sample be representative of the general data. The loss function measures a discrepancy between the output of NN applied to given arguments and their labels from the training sample. A typical NN has a network structure with layers of interconnected nodes (Fig. 2) inspired by mathematical modeling of a biological brain. Each connection has a weight. Each node has an activation function that maps the node’s input to the node’s output. The node’s input can be defined as a scalar product of the vector of outputs of the nodes connected to this node and the vector of weights of the corresponding edges plus an intercept. The weights and intercepts are estimated by the supervised learning. Neural networks were successfully applied for image recognition and recently were extended to other fields (Skansi, 2018). For more information about neural networks and machine learning terminology used in this paper, the reader is referred to Goodfellow et al. (2016) or Skansi (2018).

Fig. 2 Structure of the NN

Before describing the NN, a specific transformation is built based on the assumption that the ICC is discrete and monotonically non-decreasing. This one-to-one transformation maps the discrete ICC to the probability mass function (PMF), where $n = 7$ is the number of ability levels $(-3, -2, -1, 0, 1, 2, 3)$ indexed as $(1, 2, 3, 4, 5, 6, 7)$. Direct mapping is used to create labels, and inverse mapping is used to predict ICCs; they are defined by the following two equations, respectively:

$$\begin{aligned}
 \text{PMF}[1] &= \text{ICC}[1] \\
 \text{PMF}[i] &= \text{ICC}[i] - \text{ICC}[i - 1], \quad i = 2, 3, \dots, n \\
 \text{PMF}[n + 1] &= 1 - \text{ICC}[n]
 \end{aligned} \tag{1}$$

$$\begin{aligned}
 \text{ICC}[1] &= \text{PMF}[1] \\
 \text{ICC}[i] &= \text{PMF}[i] + \text{PMF}[i - 1], \quad i = 2, 3, \dots, n
 \end{aligned} \tag{2}$$

There are numerous degrees of freedom in terms of the number of layers, the types of layers, the number of nodes in each layer, the types of activation functions, and the types of regularizations, all affecting properties of the corresponding NN. These so-called hyperparameters are usually identified via a cross-validation study, which is discussed later in this section. The result of that study is the following NN (Fig. 2):

1. Input layer with nodes corresponding to features extracted directly from an item (see the next section about the actual features used in this study), where the number of nodes corresponds to the number of features used for predicting the ICC.
2. Layer with 16 nodes and ReLU activation function $f(x) = \max(0, x)$ for each node.

3. Dropout regularization layer with 0.1 rate that functions as follows: during the training of the NN, the output of a node from the previous layer (Fig. 2) is dropped with probability 0.1. This layer prevents an overfitting of the NN to the training data.
4. Output layer with softmax activation function $f(z_i) = \exp(z_i)/S$, $S = (\exp(z_1) + \exp(z_2) + \dots + \exp(z_8))$, $i = 1, 2, \dots, 8$, where z_i is input of node i of the output layer. This layer allows the NN to perform a soft classification, where each class gets a positive probability of being assigned such that the sum of probabilities is 1. This is in contrast to a conventional classification, where the class assignment probabilities are from $\{0, 1\}$. Thus, the output of the NN is the PMF.

A common choice of loss function for the soft classification is cross entropy, since (usually in practice) one class is intended to be selected. However, in this study, the PMF (corresponding to the ICC) should be matched as closely as possible; therefore, Kullback-Leibler divergence (Kullback & Leibler, 1951) was chosen as a loss function. This technique also works as an additional regularization to ensure that predicted ICCs are always monotonically non-decreasing.

In order to train the NN, a stochastic gradient descent (SGD, Goodfellow et al., 2016) minimized the loss function under the following parameters: learning rate 0.01; number of epochs 30; number of samples per gradient update 100. The parameters of SGD, the number of layers in the NN, the number of hidden layers (i.e., layers between the first layer and the last layer), the number of nodes in hidden layers, and the drop rate of 0.1 for the regularization layer were chosen during multiple empirical trials in order to achieve a stable output of a cross-validation described next. The NN has only one hidden layer with 16 nodes (Fig. 2); any increase either in the number of hidden layers or in the number of nodes degraded the results.

To validate the NN, the k -fold cross-validation method was used (Goodfellow et al., 2016), where labeled data is divided into k non-overlapping samples. Then, in each iteration (out of k iterations total), the $k - 1$ samples are used to train the NN and 1 sample is used to test the NN on predicting ICCs. This way, each data point is used once to train the NN and once to test it. In this study, $k = 10$ in order to comply with studies reviewed by Ferrara et al. (2021). The output of the validation includes error (E), residual (R), and coefficient of determination D computed using errors on true and average (estimated on testing samples) ICCs.

With a true ICC (given the ICC of an item) and its prediction by the NN, denoted as ICC and ICC*, respectively, the error is computed as follows:

$$E = \left((ICC [1] - ICC^* [1])^2 + \dots + (ICC [7] - ICC^* [7])^2 \right) / 7 \tag{3}$$

And the residual is computed as follows:

$$R = \left((ICC [1] - ICC^* [1]) + \dots + (ICC [7] - ICC^* [7]) \right) / 7 \tag{4}$$

Additional output includes outliers with E over 0.15 and data points with R within a certain range.

3 Data

A total of 1742 retired logical reasoning (LR) items from the Law School Admission Test were used in this study to build a prediction model. Each item has the following structure:

1. Passage
2. Question
3. Five answer options (A, B, C, D, E)

Detailed information about each LR item can be described as follows:

1. Text of the passage
2. Text of the question
3. Text of each answer option (A, B, C, D, E)
4. Correct option index (the key)
5. Item type
6. Item subtype defining the type of question
7. Item property 1
8. Item property 2
9. Item property 3
10. Item property 4
11. Item rank {1, 2, 3, 4} (an estimation of item difficulty by item writer)
12. Item pretest position (item position, defined by item writer, in unscored section for pretesting new items)
13. ICC computed from corresponding 3PLM (response matrices were not available for computing empirical ICCs)

The above information was used to compute multiple features for each item. An additional 2009 LR items without item rank and item pretest position were used to compute some numerical features for the above 1742 LR items. As a rule of thumb, an acceptable performance of an NN is observed when there are around 5000 labeled data points per category (Goodfellow et al., 2016). Therefore, the studied data is too small to expect superior results; in fact, because there are nine different question types, the data is partitioned into even smaller pieces.

The rest of this section describes features extracted directly from an item. For each item, six categorical features were provided by the item writer:

1. Item type (categorical feature from {1, 2})
2. Item subtype (categorical feature from {1, 2, 3, 4, 5, 6, 7, 8, 9})
3. Item property 1 (categorical feature from {1, 2, 3})
4. Item property 2 (categorical feature from {1, 2, 3, 4, 5})

5. Item property 3 (categorical feature from {1, 2, 3})
6. Item property 4 (categorical feature from {1, 2})

Each categorical feature was represented as a one-hot code vector (Goodfellow et al., 2016); for example, if the item type (see above) was 1 then it was represented as vector (1, 0), and if the item type was 2 then it was represented as vector (0, 1). Using this representation allows neural networks to apply the divide-and-conquer strategy similarly to CARTs (Breiman et al., 1984). There were 12 groups of numerical features:

1. Item rank, denoted as *itemRank*.
2. Item pretest position, denoted as *itemPosition*.
3. Text features for passage: *PTF.nSentences* (number of sentences), *PTF.nWords* (number of words), *PTF.nNouns* (number of nouns), *PTF.nNSynsets* (number of synsets (Fellbaum, 1998) of the nouns), *PTF.nVerbs* (number of verbs), *PTF.nVSynsets* (number of synsets of the verbs), *PTF.nAdjs* (number of adjectives), *PTF.nASynsets* (number of synsets of the adjectives), *PTF.readability* (Dale–Chall readability index; Dale & Chall, 1948).
4. Text features for question: *QTF.nWords* (number of words), *QTF.nNouns* (number of nouns), *QTF.nNSynsets* (number of synsets of the nouns), *QTF.nVerbs* (number of verbs), *QTF.nVSynsets* (number of synsets of the verbs), *QTF.nAdjs* (number of adjectives), *QTF.nASynsets* (number of synsets of the adjectives), *QTF.readability* (Dale–Chall readability index).
5. Text features for options: *OTF.nSentences* (number of sentences), *OTF.nWords* (number of words), *OTF.nNouns* (number of nouns), *OTF.nNSynsets* (number of synsets of the nouns), *OTF.nVerbs* (number of verbs), *OTF.nVSynsets* (number of synsets of the verbs), *OTF.nAdjs* (number of adjectives), *OTF.nASynsets* (number of synsets of the adjectives), *OTF.readability* (Dale–Chall readability index).
6. Semantic similarity between passage and correct option (answer) denoted as *s_{pa}*. Semantic similarity between two texts is computed as a scalar product between two embeddings corresponding to two texts; for information about embeddings see Goodfellow et al. (2016).
7. Semantic similarity between passage with correct option (answer) and other options (distractors), denoted as *s_{pad}*.
8. Mean, variance, minimum value, and maximum value of semantic similarity between answer and distractors, denoted as: *sadMean*, *sadVar*, *sadMin*, *sadMax*.
9. Mean, variance, minimum value, and maximum value of semantic similarity between all unique pairs of options, denoted as: *sooMean*, *sooVar*, *sooMin*, *sooMax*.
10. Mean, variance, minimum value, and maximum value of semantic similarity between all unique pairs of sentences in the passage, denoted as: *sppMean*, *sppVar*, *sppMin*, *sppMax*.
11. The additional 2009 LR items (called *atlas items*) without item rank and item pretest position were used to compute this group of features. The atlas

items were partitioned into nonintersecting classes based on their type and subtype. An item, used for constructing the NN, was associated with a class corresponding to the type and subtype of the item. Each element of the associated class had the passage and difficulty of some atlas item. For each element in the class, the semantic similarity between the element’s passage and the item’s passage multiplied by the element’s difficulty was sampled. Finally, the mean, variance, minimum value, and maximum value were estimated from the sample and denoted as: *sppbMean*, *sppbVar*, *sppbMin*, *sppbMax*.

12. Similarly to the previous group, the following features were computed for the question: *sqqbMean*, *sqqbVar*, *sqqbMin*, *sqqbMax*.

Each numerical feature from above was normalized by subtracting its mean and then dividing by its standard deviation; such normalization substantially improves the convergence of SGD (Goodfellow et al., 2016).

Table 1 shows features that correlate with at least $|0.1|$ with *a*, *b*, or *c* parameters of 3PLM. One may observe some interesting patterns in Table 1. The highest correlations are observed for *itemRank* and *itemPosition*. Most correlations are with item difficulty except for *sppbMean* and *sqqbMean* (which perhaps relate to their estimation procedure). Most text features in Table 1 are for options, only two for

Table 1 Features that correlate with at least $|0.1|$ with *a*, *b*, or *c* parameters of 3PLM

Feature (see description of each feature above)	Correlation with <i>a</i>	Correlation with <i>b</i>	Correlation with <i>c</i>
<i>itemRank</i>	0.10	0.43	0.07
<i>itemPosition</i>	0.12	0.43	0.08
<i>sppbMean</i>	0.12	0.15	0.17
<i>sqqbMean</i>	0.14	0.17	0.21
<i>sqqbVar</i>	0.00	0.12	0.02
<i>sqqbMax</i>	0.02	0.12	0.03
<i>sppMean</i>	0.06	0.11	-0.02
<i>sadMean</i>	0.07	0.10	0.03
<i>sadMin</i>	0.06	0.11	0.03
<i>sooMean</i>	0.07	0.11	0.01
<i>sooMin</i>	0.07	0.11	0.03
<i>QTF.nVerbs</i>	0.10	0.08	0.07
<i>QTF.nVSynsets</i>	0.13	0.10	0.07
<i>OTF.nSentences</i>	0.06	0.15	0.11
<i>OTF.nWords</i>	0.05	0.20	0.12
<i>OTF.nNouns</i>	0.01	0.14	0.07
<i>OTF.nNSynsets</i>	0.01	0.13	0.03
<i>OTF.nVerbs</i>	0.04	0.18	0.14
<i>OTF.nVSynsets</i>	0.02	0.15	0.11
<i>OTF.nAdjs</i>	0.03	0.12	0.03
<i>OTF.readability</i>	-0.07	-0.12	-0.04

question, and none for the passage, which is unexpected. Overall, text features have higher absolute correlations than features based on semantic similarity. Feature *sppMean* has a positive correlation with item difficulty, which means that more difficult items may have more closely related sentences in their passages (this is supported by real data).

4 Results

Table 2 shows the results of 10-fold cross-validation of the NN for different subsets of features. One can observe that the best results were achieved with features provided only by item writers (see fifth column in Table 2). Thus, the use of automatically generated features did not improve the results, although some of them weakly correlate with *a*, *b*, or *c* parameters of 3PLM (Table 1).

The cross-validation of the NN constructed from categorical features and numerical features *itemRank*, *itemPosition* provided additional results as follows. Graphical representation of error and residual computed for each ability level separately is illustrated in Figs. 3 and 4, where distributions of error and residual are characterized by box plots. One can see that the largest errors and residuals happened for ability level 0. Figure 5 shows a random sample of nine pairs of true and predicted ICCs, shown as blue and green curves, respectively, where the residual fell within one standard deviation from its mean; overall, 67% of predicted ICCs satisfied that range. One can observe that true and predicted ICCs are different in terms of variability (see Fig. 5), and that the variability of true ICCs is higher than the variability of predicted ICCs (Fig. 6; this finding is compatible with low values of *D* in Table 2).

5 Discussion

This paper describes the NN approach to predicting ICCs using features extracted directly from an item. A total of 1742 retired LR items from the LSAT were used to build, train, and validate the NN.

Multiple features extracted directly from an item were used in the input layer of the NN (see Fig. 2). A cross-validation study using different subsets of the features demonstrated (see Table 2) that using features provided by item writers (categorical features and numerical features *itemRank*, *itemPosition*) produced the best predictions whereas automatically generated features did not improve the predictions. Even more, just using two features (*itemRank* and *itemPosition*) produced the second best results. This indicates that the data sample is too small for the categorical features to play any role in prediction. That may also explain why automatically generated features were useless, since the number of items in each

Table 2 Results of 10-fold cross-validation of the NN for different subsets of features

Measure	Categorical and numerical features	Categorical and numerical features from Table 1	Categorical and numerical features from Table 1 without <i>itemRank</i> , <i>itemPosition</i>	Categorical and numerical features from <i>itemRank</i> , <i>itemPosition</i>	Only <i>itemRank</i> and <i>itemPosition</i>
<i>E</i> mean	0.022	0.021	0.024	0.020	0.021
<i>E</i> st.dev.	0.025	0.023	0.024	0.022	0.022
<i>R</i> mean	-0.003	0.000	0.000	-0.001	0.000
<i>R</i> st.dev.	0.120	0.116	0.125	0.115	0.116
<i>R</i> skewness	-0.188	-0.205	-0.150	-0.194	-0.191
<i>D</i>	0.105	0.153	-0.007	0.178	0.160
Number of outliers	8	2	1	1	0
Min <i>E</i> on the outliers	0.151	0.190	0.153	0.161	-
Max <i>E</i> on the outliers	0.239	0.192	0.153	0.161	-

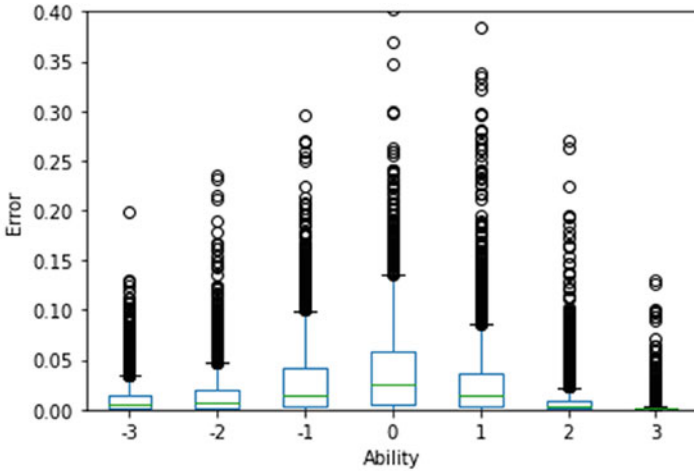


Fig. 3 Box plots of error for each ability level

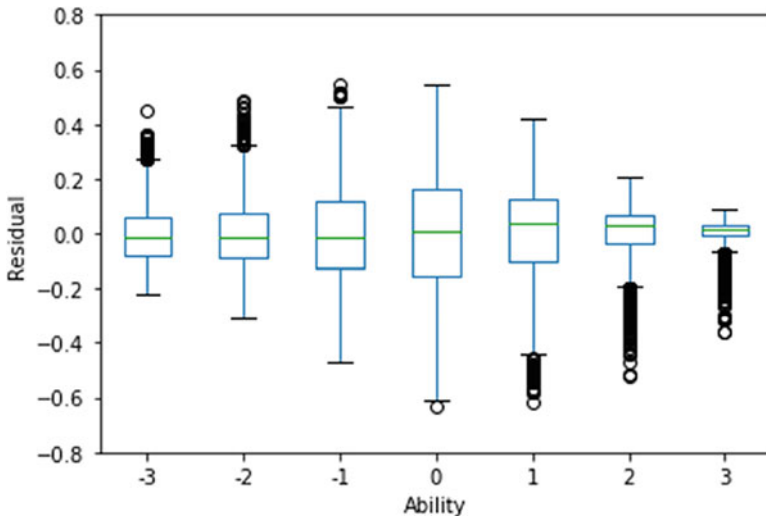


Fig. 4 Box plots of residual for each ability level

category was even smaller. The best subset of features provided a low D , which was expected as the data was too small.

This study is in line with 90% of the studies (Ferrara et al., 2021) reporting a low coefficient of determination D . This paper confirms that the best predictors are features provided by item writers (see Table 2). As expected, in contrast to reading comprehension items, LR items have a weak correlation between text features and item difficulty (see Table 1).

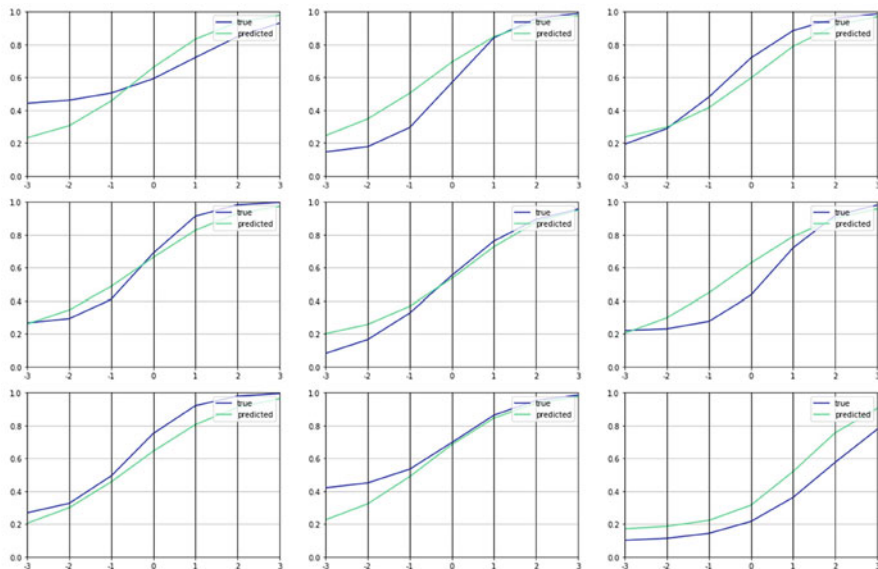


Fig. 5 A random sample of nine pairs of true and predicted ICCs (blue and green curves respectively), where the residual was within one standard deviation from its mean. Overall, 67% of predicted ICCs fell within that range

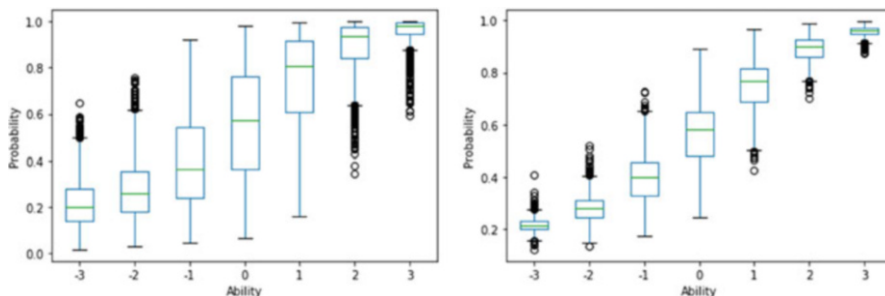


Fig. 6 Distribution of true ICCs (left box plot) and predicted ICCs (right box plot)

Online test proctoring cannot protect against existing technology used to steal test content. Therefore, one has to accept an error generated by a predictive model if the error is symmetrically distributed about zero. A possible application scenario would be as follows: (1) develop a model predicting ICC; (2) use predicted ICCs to simulate a real administration of pretested sections; (3) based on simulated responses, calibrate items (e.g., calibrate 3PLMs); (4) assemble a test using these items as already pretested; (5) administer the test to a real population; (6) use real responses for recalibrating items and updating the model. In this scenario, new items are not pretested (i.e., not administered to a real subpopulation of examinees). Therefore, a regular scaling based on anchor items pretested in the past is no longer

possible. This can be addressed by using an additional, previously administered section as an anchor section. Modern methods of detecting item preknowledge (Belov, 2016, 2020; Drasgow et al., 1996; Karabatsos, 2003; McLeod et al., 2003; Tendeiro & Meijer, 2012; van Krimpen-Stoop & Meijer, 2001) can be applied to filter out examinees with possible preknowledge of the anchor section in order to estimate scaling coefficients without bias.

If presented approach were applied to predict empirical ICCs, then the assumption of monotonically non-decreasing ICCs could be violated by some empirical ICCs (Fig. 1). In this case, the NN could be modified as follows: The output layer with linear activation function could have seven nodes corresponding to ability levels ($-3, -2, -1, 0, 1, 2, 3$), and the loss function could be the mean squared error.

The approach could be easily adapted to predict parameters of 3PLM directly. The only modification would be that the output layer with linear activation function would have three nodes (for a, b , and c , respectively) and the loss function would be the mean squared error.

Future research will be directed toward procuring a larger data sample, engineering new features, minimizing E , and maximizing D , while keeping R symmetrically distributed about zero. The latter is crucial in order for items with predicted statistical parameters to be included on a test. Larger data may allow the use of embeddings (Goodfellow et al., 2016) for the passage, question, and options directly (instead of computing features as semantic similarities between various parts of the item, as was done in this study); that way a deeper NN could “figure out” more useful features. Text features used in this study can be extended with Coh-Metrix (Ferrara et al., 2021). Another method to generate new features is described in the final two groups of numerical features in the Data section above. For a given item, the method could be generalized as follows. From the atlas items (items without item rank and item pretest position), form a class using a certain criterion; for example, select items with multiple negations in their passages. Each element of the class has the passage and difficulty of some atlas item. For each element in the class, the semantic similarity between the element’s passage and the item’s passage, multiplied by the element’s difficulty, is sampled. Then the mean, variance, minimum value, and maximum value estimated on the sample could be used as the new features.

References

- Belov, D. I. (2016). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement, 40*, 83–97.
- Belov, D. I. (2020). Monte Carlo detection of examinees with item preknowledge. *Behav-iormetrika, 48*, 23–50.
- Breiman, L., Friedman, J. H., Olshen, R., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth International Group.

- Dale, E., & Chall, J. (1948). A formula for predicting readability. *Educational Research Bulletin*, 27, 11–20.
- Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education*, 9, 47–64.
- Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. MIT Press.
- Ferrara, S., Steedle, J., & Frantz, R. (2021). *Response demands of reading comprehension items: A review of item difficulty modeling studies* [Manuscript submitted for publication]
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- Huff, K. (2006). *Using item difficulty modeling to inform descriptive score report*. Annual Meeting of the National Council on Measurement in Education, San Francisco, CA, United States
- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16, 277–298.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22, 79–86.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum.
- McLeod, L. D., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, 27, 121–137.
- Sheehan, K., & Mislavy, R. J. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement*, 27(3), 255–272.
- Skansi, S. (2018). *Introduction to deep learning*. Springer.
- Tendeiro, J. N., & Meijer, R. R. (2012). A CUSUM to detect person misfit: A discussion and some alternatives for existing procedures. *Applied Psychological Measurement*, 36, 420–442.
- van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics*, 26, 199–217.

Pooled Autoregressive Models for Categorical Data



Zhenqiu (Laura) Lu and Zhiyong (Johnny) Zhang

Abstract Time series capture time dependent intra-individual variation within a single participant. When data are collected from more than one subject, methods developed for single subject intra-individual relationship may not fully work and laws governing inter-individual relationship may not apply to intra-individual relationship, especially when outcomes are categorical or ordinal data. These data are usually collected by the Likert table. This article aims to investigate the performance of four estimation methods for pooling time series data focusing on categorical outcomes and to address related issues through an autoregressive model, AR(1). In this article, models for pooling time series were formulated, estimation methods were derived, simulation studies were conducted, results were summarized and compared.

Keywords Pooling time series · Autoregressive model · Categorical data · Conditional likelihood · Exact likelihood · Maximum likelihood estimation

1 Introduction

The variation analysis in psychological, social, and behavioral researches has many ramifications. Among them two main branches are inter-individual variation and intra-individual variation. Inter-individual variation is the variation between individuals, and also widely known as the analysis of cross-sectional data in many researches. Intra-individual variation is the time dependent variation within a single participant's time series. It is also known as the analysis of time series data or P-technique in Cattell's (1952) data-box Cattell (1952). In this type of study, usually

Z. (Laura) Lu (✉)
University of Georgia, Athens, GA, USA
e-mail: zlu@uga.edu

Z. (Johnny) Zhang
University of Notre Dame, Notre Dame, IN, USA

one subject is measured and the variables of interests are collected from each of a large number of occasions. Many methods are available for single time series analysis (e.g., Cattell et al., 1947; Molenaar, 1985; Nesselroade & Molenaar, 2003).

However, data collected in this way do not have inter-individual differences since there is only one subject involved, but they can reflect changes across occasions. Intra-individual analysis has become popular advanced by Nesselroade, Molenaar, and colleagues. So many researches on intra-individual relationship, data are collected from more than one subject. When multiple subjects are involved, methods developed for single subject intra-individual relationship may not fully work. Also, laws governing inter-individual relationship may not apply to intra-individual relationship (Molenaar, 2004; Nesselroade & Ram, 2004, e.g.). There are few methods in literature dealing with the analysis of pooling multiple time series (Cattell & Scheier, 1961; Daly et al., 1974; Molenaar et al., 2003; Nesselroade & Molenaar, 1999, e.g.). The attention of this article will be drawn to multiple subjects intra-individual variation analysis. Also, the data in educational and social areas are usually collected by Likert tables. But the research on multiple subjects time series for categorical outcomes is very few. So we fill the gap by focusing our research in this area.

This article aims to investigate the performance of different estimation methods for pooling time series data focusing on categorical outcomes and to address related issues through an AR(1) model. We focus on four estimation methods for multiple time series: pooling conditional likelihood estimation, pooling exact likelihood estimation, connecting data conditional likelihood, and connecting data exact likelihood.

This article is organized as follows. In the next section some introductory remarks about time series are given. First single series and multiple series focusing on the AR(1) model are described and formulated. And then different estimation methods for multiple time series are introduced and derived. Then follows a section of simulation studies in which the performance of four estimating methods are investigated under various conditions. Simulation results are provided after simulation design and implementation. The closing part of this article summarizes the simulation results, compares different estimation methods of aggregating time series, and provides practical implication.

2 Autoregressive Model and Categorical Data

2.1 First-Order Autoregressive Model, AR(1)

We first consider a model for a single subject (or individual). And then we extend it to the model for multiple subjects. Suppose we are interested in a first-order autoregressive model, AR(1), as follows.

y_1 : the initial value

$$y_t = \mu + \alpha y_{t-1} + z_t \quad (t > 1) \quad \text{with} \quad z_t \sim i.i.d. N(0, \phi) \quad (1)$$

where y_t is the observed value at time point t , α is the model autoregressive coefficient, μ is a parameter correlated with the mean of y , z is a shock variable, or a white noise sequence, satisfying a normal distribution with mean 0 and variance ϕ . In this case, the vector of population parameters to be estimated consists of $\theta = (\mu, \alpha, \phi)'$. When $|\alpha| < 1$, there is a covariance stationary process for y_t satisfying Eq (1). Thus, the remainder of this discussion of AR(1) assumes that $|\alpha| < 1$. By algebra and Taylor expansion, we have the mean, the variance, and the j^{th} autocovariance of y_t .

$$E(y_t) = \frac{\mu}{1 - \alpha}, \quad (2)$$

$$\text{Var}(y_t) = \frac{\phi}{1 - \alpha^2}, \quad (3)$$

$$\text{Cov}(y_t, y_{t-j}) = \alpha^j \frac{\phi}{1 - \alpha^2} \quad (4)$$

So we have the following distribution of y_t

$$\begin{cases} y_1 \sim N\left(\frac{\mu}{1-\alpha}, \frac{\phi}{1-\alpha^2}\right), \\ y_t | y_{t-1} \sim N(\mu + \alpha y_{t-1}, \phi), \quad (t > 1). \end{cases}$$

For multiple subjects, suppose there are N individuals, we can express the constant coefficient AR(1) model as follows:

$$y_{it} = \mu + \alpha y_{i(t-1)} + z_{it}, \quad (i = 1, \dots, N; t = 2, \dots, T)$$

where $z_{it} \ i.i.d. \sim N(0, \phi)$ and the parameters μ , α and ϕ are constants which keep the same values across all individuals. This model is very useful when the sample size (or number of participants) is small but with fairly large measurement occasions.

2.2 Categorical Data

Let c be the number of categories and $\tau = (\tau_1, \dots, \tau_{c-1})$ be thresholds. Assume y_{it} is a continuous normality distributed variable following AR(1) model $y_{it} = \mu + \alpha y_{i(t-1)} + z_{it}$. With the assumption of the normality distribution of y_t , the thresholds τ can be created from the standardized thresholds τ_z as

$$\tau = \mu_y + \tau_z \sigma_y$$

where $\mu_y = \frac{\mu}{1-\alpha}$ and $\sigma_y = \sqrt{\frac{\phi}{1-\alpha^2}}$. With thresholds τ , categorical data y_{it}^* can be created by

$$\begin{cases} y_{it}^* = 1, & \text{when } y_{it} \leq \tau_1; \\ y_{it}^* = k, & \text{when } \tau_{k-1} < y_{it} \leq \tau_k; \\ y_{it}^* = c, & \text{when } y_{it} > \tau_{c-1}. \end{cases}$$

The scale of y_{it}^* is from 1 to c . Let π be a c -dimensional vector $\pi = (\pi_1, \dots, \pi_c)$ which is defined as

$$\begin{aligned} \pi_1 &= \Phi(\tau_1) \\ \pi_k &= \Phi(\tau_k) - \Phi(\tau_{k-1}) \quad (2 \leq k \leq c-1) \\ \pi_c &= 1 - \Phi(\tau_{c-1}), \end{aligned}$$

then each π_k ($1 \leq k \leq c$) is defined to be the probability of corresponding k th category. With π , the mean of c categories is

$$M_c = \sum_{k=1}^c \pi_k k$$

Therefore, the true μ and ϕ of c categories are

$$\begin{aligned} \mu_c &= M_c (1 - \alpha), \\ \phi_c &= \left[\sum_{k=1}^c \pi_k (k - M_c)^2 \right] (1 - \alpha^2), \end{aligned}$$

3 Estimation Methods and Likelihoods

This study investigates two MLE estimation methods: (1) exact MLE estimation method: the parameters are estimated by maximizing the exact log-likelihood function including the distribution of deterministic y_1 which requires stationarity assumption, and (2) conditional MLE estimation method: the parameters are estimated by maximizing the conditional log-likelihood function without y_1 . For multiple subjects we pool likelihood functions for all individuals. In practice, there is another method to deal with time series data by connecting all similar time series from multiple subjects together as from a single subject (reference here). It assumes there is some relationship between y_{iT} and $y_{(i+1)1}$. We have pooled data exact MLE and pooled data conditional MLE.

3.1 Exact MLE for Pooled Likelihood Function

The exact likelihood function of the stationary AR(1) model described in Eq. (1) and its corresponding log likelihood function are

$$\begin{aligned}
 L_i(\alpha, \mu, \phi | \mathbf{y}_i) &= \frac{1}{\sqrt{2\pi(\frac{\phi}{1-\alpha^2})}} \exp\left[-\frac{(y_{i1} - \frac{\mu}{1-\alpha})^2}{2(\frac{\phi}{1-\alpha^2})}\right] \\
 &\quad \times \left\{ \prod_{t=2}^T \frac{1}{\sqrt{2\pi\phi}} \exp\left[-\frac{(y_{it} - \mu - \alpha y_{i(t-1)})^2}{2\phi}\right] \right\}, \\
 \log(L) &= \frac{N}{2} \log(1 - \alpha^2) - \frac{1 - \alpha^2}{2\phi} \sum_{i=1}^N (y_{i1} - \frac{\mu}{1 - \alpha})^2 - \frac{NT}{2} \log(2\pi\phi) \\
 &\quad - \frac{1}{2\phi} \sum_{i=1}^N \sum_{t=2}^T (y_{it} - \mu - \alpha y_{i(t-1)})^2.
 \end{aligned}$$

In order to obtain the maximum likelihood estimates (MLE) of parameters μ , α and ϕ , we make all of their first order derivatives with respect to these parameters 0 and their corresponding second order derivatives negative. The MLE obtained through solving the exact likelihood function is called the exact MLE. Unfortunately, there is no simple solution for θ in terms of $(\{y_{it}\}, 1 \leq i \leq N, 1 \leq t \leq T)$. But with the help of computers, we can use iterative or numerical procedures to solve the equation.

3.2 Exact MLE for Pooled Data

The exact likelihood function of the connected stationary AR(1) model and its corresponding log likelihood function are

$$\begin{aligned}
 L(\alpha, \mu, \phi | \mathbf{y}) &= \frac{1}{\sqrt{2\pi(\frac{\phi}{1-\alpha^2})}} \exp\left[-\frac{(y_1 - \frac{\mu}{1-\alpha})^2}{2(\frac{\phi}{1-\alpha^2})}\right] \\
 &\quad \times \left\{ \prod_{t=2}^{NT} \frac{1}{\sqrt{2\pi\phi}} \exp\left[-\frac{(y_t - \mu - \alpha y_{t-1})^2}{2\phi}\right] \right\}, \\
 \log(L) &= \frac{1}{2} \log(1 - \alpha^2) - \frac{1 - \alpha^2}{2\phi} (y_1 - \frac{\mu}{1 - \alpha})^2 \\
 &\quad - \frac{NT}{2} \log(2\pi\phi) - \frac{1}{2\phi} \sum_{t=2}^{NT} (y_t - \mu - \alpha y_{t-1})^2.
 \end{aligned}$$

By making the first derivatives zero equal to 0 to obtain the solution. Again, unfortunately, there is no simple solution for θ in terms of $(\{y_{it}\}, 1 \leq i \leq N, 1 \leq t \leq T)$.

3.3 Conditional MLE for Pooled Likelihood Function

The conditional likelihood function of the stationary AR(1) model does not take the distribution of y_1 into consideration, so the likelihood and its log likelihood function are

$$L_i(\alpha, \mu, \phi | \mathbf{y}_i) = \prod_{t=2}^T \frac{1}{\sqrt{2\pi\phi}} \exp \left[-\frac{(y_{it} - \mu - \alpha y_{i(t-1)})^2}{2\phi} \right], \quad (5)$$

$$\log(L(\alpha, \mu, \phi | \mathbf{y})) = -\frac{N(T-1)}{2} \log(2\pi\phi) - \frac{1}{2\phi} \sum_{i=1}^N \sum_{t=2}^T (y_{it} - \mu - \alpha y_{i(t-1)})^2. \quad (6)$$

To obtain the MLE of parameters μ , α and ϕ , we make their first derivatives zero and the second order derivatives negative, and we have

$$\hat{\mu} = \frac{1}{N(T-1)} \sum_{i=1}^N \sum_{t=2}^T (y_{it} - \hat{\alpha} y_{i(t-1)}) \quad (7)$$

$$\hat{\alpha} = \frac{\sum_{i=1}^N \sum_{t=2}^T [(y_{it} - \hat{\mu}) y_{i(t-1)}]}{\sum_{i=1}^N \sum_{t=2}^T y_{i(t-1)}^2} \quad (8)$$

$$\hat{\phi} = \frac{1}{N(T-1)} \sum_{i=1}^N \sum_{t=2}^T (y_{it} - \hat{\mu} - \hat{\alpha} y_{i(t-1)})^2 \quad (9)$$

We can also use the ordinal least square (OLS) estimation method to obtain μ and α ,

$$\hat{\beta} = \begin{bmatrix} \hat{\mu} \\ \hat{\alpha} \end{bmatrix} = \begin{bmatrix} N(T-1) & \sum_{i=1}^N \sum_{t=2}^T y_{i(t-1)} \\ \sum_{i=1}^N \sum_{t=2}^T y_{i(t-1)} & \sum_{i=1}^N \sum_{t=2}^T y_{i(t-1)}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^N \sum_{t=2}^T y_{it} \\ \sum_{i=1}^N \sum_{t=2}^T y_{i(t-1)} y_{it} \end{bmatrix}.$$

And ϕ can be obtained by inserting the estimates of (μ, α) into Eq. (9). The OLS solution is exactly the same as the MLE solution.

3.4 Conditional MLE for Pooled Data

The conditional likelihood function and its corresponding log likelihood function of the connected stationary AR(1) model are

$$L(\alpha, \mu, \phi|\mathbf{y}) = \prod_{t=2}^{NT} \frac{1}{\sqrt{2\pi\phi}} \exp \left[-\frac{(y_t - \mu - \alpha y_{t-1})^2}{2\phi} \right],$$

$$\log(L) = -\frac{NT - 1}{2} \log(2\pi\phi) - \frac{1}{2\phi} \sum_{t=2}^{NT} (y_t - \mu - \alpha y_{t-1})^2.$$

4 Simulation Study

We conduct a simulation study to investigate the performance of the exact MLE estimation method and conditional MLE estimation method fitting different models fitting categorical data. We use iterative or numerical procedures to solve the equations which have no explicit solutions.

4.1 Data Generation

The true values in this simulation are set as $\mu = \mu_c, \alpha = 0.5, \phi = \phi_c$. The replication number is 1000. We use the following 3 steps to generate the categorical data y_{it}^* .

Step 1: Generate the continuous data according to the constant coefficients AR(1) model $y_{it} = \mu + \alpha y_{i(t-1)} + z_{it}$.

Step 2: Generate thresholds $\tau = (\tau_1, \dots, \tau_{c-1})$. With the assumption of the normality distribution of y_{it} , the thresholds τ are created by (1) obtaining the standardized thresholds τ_z by dividing the segment $[-2, 2]$ into $c - 2$ parts evenly, and then (2) transform τ_z to τ according to the original data scale. For example, if $c = 5$, the standardized thresholds are $\tau_z = (\tau_{z1}, \tau_{z2}, \tau_{z3}, \tau_{z4}) = (-2, -2/3, 2/3, 2)$, then $\tau = \mu_y + \tau_z \sigma_y$.

Step 3: Generate the categorical data y_{it}^* by

$$\begin{cases} y_{it}^* = 1, & \text{when } y_{it} \leq \tau_1; \\ y_{it}^* = k, & \text{when } \tau_{k-1} < y_{it} \leq \tau_k; \\ y_{it}^* = c, & \text{when } y_{it} > \tau_{c-1}. \end{cases}$$

Simulation condition factors in this study include the initial value, the number of categories, the lengths of series, and the number of subjects. (1) The initial value

y_1 has 3 cases: (i) a fixed yc_1 based on a fixed $y_1 = 0$; (ii) a random yc_1 based on a random y_1 from $N(0, \phi)$; and (iii) a random yc_1 based on a random y_1 from $N(\frac{\mu}{1-\alpha}, \frac{\phi}{1-\alpha^2})$. (2) The number of categories is $c = (5, 7, 9)$. (3) The lengths of series is set as $T = (5, 10, 15, 20, 30, 40, 50)$ to catch the change patterns. (4) The number of subjects is $N = (50, 100, 150, 200)$. In total, there are $3 * 3 * 7 * 4 = 252$ conditions, with each condition having 1000 replications.

4.2 Model Estimation and Evaluation

When the categorical data are ready, we use four estimation methods: pooled likelihood exact MLE, pooled likelihood conditional MLE, pooled data exact MLE, and pooled data conditional MLE. We use MSE, the mean square error of the estimate, to compare accuracy of estimates.

$$MSE = Bias.abs^2 + SE.emp^2$$

where $Bias.abs$ is the absolute bias of the estimate, and $SE.emp$ is the empirical standard error across 1000 replications.

R language was used to generate data, estimate parameters, and summarize results. The main R functions for data generating and model estimation are attached in Appendix 1.

5 Results, Conclusions and Discussion

5.1 Results

In total, there are 252 (= 3 initial values \times 3 numbers of categories \times 7 lengths of series \times 4 number of subjects) simulation conditions. For each condition, there are 4 estimation methods. Part of simulation results are summarized and shown in Tables 1, 2, and 3. For example, Table 1 summarized part of the estimation results from 1000 replications with $c = 5$ categories, including sample size $N = 50$ or $N = 200$ individuals, time series length $T = 5$ or $T = 50$ observations per individual, and the initial value of y from $y_1 = 0$, $y_1 \sim N(0, \psi)$ or $y_1 \sim N(\frac{\mu}{1-\alpha}, \frac{\psi}{1-\alpha^2})$.

From Table 1, we see that the max value of MSE under all conditions of $N = 200$ and $T = 50$ is 0.0516 and the min value is 0.0009. But under the conditions with fewer individual participated $N = 50$ and shorter time series $T = 5$, the max and min values of MSE are 0.3596 and 0.0021, respectively. The smaller MSE value, the more accurate the estimate. So the longer the time series or the more individual participated, the more accurate the estimate. By comparing all three Tables Tables 1, 2, and 3 with difference categories, we can also see

Table 1 Results summarized from 1000 replications for $c = 5$ categories

		N = 50										N = 200									
		True ^a		Est. ^b	Bias.abs ^c	Bias.rel ^d	SE.emp ^e	SE.avgf	MSE ^g	Cover ^h	Est.	Bias.abs	Bias.rel	SE.emp	SE.avg	MSE					
$y_1 = 0$																					
P.L. ⁱ	Exact ^j	μ	1.5	2.0002	0.5002	0.3335	0.1847	0.1839	0.2844	0.2180	1.7245	0.2245	0.1496	0.0289	0.0279	0.0512					
		α	0.5	0.3333	0.1667	0.3334	0.0602	0.0600	0.0314	0.2070	0.4251	0.0749	0.1497	0.0094	0.0090	0.0057					
		ψ	0.4811	0.4136	0.0675	0.1404	0.0404	0.0372	0.0062	0.5250	0.5098	0.0287	0.0597	0.0068	0.0072	0.0009					
	Cond. ^k	μ	1.5	1.7317	0.2317	0.1544	0.2360	0.2363	0.1094	0.8430	1.6985	0.1985	0.1324	0.0295	0.0286	0.0403					
		α	0.5	0.4228	0.0772	0.1544	0.0766	0.0769	0.0118	0.8370	0.4338	0.0662	0.1324	0.0096	0.0092	0.0045					
		ψ	0.4811	0.5123	0.0312	0.0648	0.0503	0.0512	0.0035	0.9390	0.5202	0.0391	0.0812	0.0069	0.0074	0.0016					
P.D. ^l	Exact	μ	1.5	2.0737	0.5737	0.3824	0.1746	0.1848	0.3596	0.1060	1.7253	0.2253	0.1502	0.0291	0.0281	0.0516					
		α	0.5	0.3090	0.1910	0.3820	0.0566	0.0600	0.0397	0.0920	0.4249	0.0751	0.1503	0.0095	0.0091	0.0057					
		ψ	0.4811	0.4280	0.0531	0.1105	0.0419	0.0383	0.0046	0.6500	0.5121	0.0310	0.0645	0.0068	0.0072	0.0010					
	Cond.	μ	1.5	2.0700	0.5700	0.3800	0.1753	0.1857	0.3556	0.1140	1.7252	0.2252	0.1501	0.0291	0.0281	0.0515					
		α	0.5	0.3102	0.1898	0.3795	0.0568	0.0603	0.0392	0.1030	0.4249	0.0751	0.1502	0.0095	0.0091	0.0057					
		ψ	0.4811	0.4297	0.0514	0.1069	0.0421	0.0385	0.0044	0.6690	0.5122	0.0311	0.0646	0.0068	0.0072	0.0010					
$y_1 \sim N(0, \psi)$																					
P.L.	Exact	μ	1.5	1.7948	0.2948	0.1966	0.1938	0.1883	0.1245	0.6750	1.7039	0.2039	0.1359	0.0287	0.0280	0.0424					
		α	0.5	0.4027	0.0973	0.1945	0.0624	0.0612	0.0134	0.6660	0.4321	0.0679	0.1358	0.0093	0.0090	0.0047					
		ψ	0.4811	0.4969	0.0157	0.0327	0.0435	0.0449	0.0021	0.9580	0.5181	0.0370	0.0769	0.0071	0.0073	0.0014					
	Cond.	μ	1.5	1.7468	0.2468	0.1645	0.2089	0.2048	0.1045	0.7900	1.6985	0.1985	0.1323	0.0290	0.0283	0.0402					
		α	0.5	0.4192	0.0808	0.1615	0.0669	0.0661	0.0110	0.7770	0.4339	0.0661	0.1322	0.0093	0.0091	0.0045					
		ψ	0.4811	0.5153	0.0342	0.0711	0.0495	0.0516	0.0036	0.9430	0.5203	0.0392	0.0814	0.0073	0.0074	0.0016					
P.D.	Exact	μ	1.5	2.0071	0.5071	0.3381	0.1864	0.1849	0.2920	0.2050	1.7247	0.2247	0.1498	0.0289	0.0281	0.0513					
		α	0.5	0.3321	0.1679	0.3358	0.0593	0.0596	0.0317	0.1830	0.4252	0.0748	0.1496	0.0093	0.0091	0.0057					
		ψ	0.4811	0.5306	0.0495	0.1029	0.0476	0.0475	0.0047	0.8620	0.5224	0.0413	0.0858	0.0072	0.0074	0.0018					
	Cond.	μ	1.5	2.0068	0.5068	0.3379	0.1867	0.1851	0.2917	0.2070	1.7247	0.2247	0.1498	0.0289	0.0281	0.0513					
		α	0.5	0.3322	0.1678	0.3355	0.0593	0.0597	0.0317	0.1880	0.4252	0.0748	0.1496	0.0093	0.0091	0.0057					
		ψ	0.4811	0.5309	0.0498	0.1034	0.0477	0.0476	0.0048	0.8630	0.5224	0.0413	0.0858	0.0072	0.0074	0.0018					

(continued)

Table 1 (continued)

		N = 50					N = 200					T = 50				
True ^a		Est. ^b	Bias.abs ^c	Bias.rel ^d	SE.emp ^e	SE.avg ^f	MSE ^g	Cover ^h	Est.	Bias.abs	Bias.rel	SE.emp	SE.avg	MSE		
$y_1 \sim N(\frac{\mu}{1-\alpha}, \frac{\psi}{1-\alpha^2})$																
P.L.	Exact	μ	1.7243	0.2243	0.1496	0.1877	0.0892	0.7930	1.6962	0.1962	0.1308	0.0285	0.0281	0.0393		
		α	0.4252	0.0748	0.1497	0.0611	0.0097	0.7900	0.4345	0.0655	0.1309	0.0093	0.0091	0.0044		
		ψ	0.4811	0.0341	0.0708	0.0466	0.0033	0.9200	0.5198	0.0387	0.0804	0.0070	0.0074	0.0015		
	Cond.	μ	1.7232	0.2232	0.1488	0.2068	0.1981	0.0926	0.8050	1.6961	0.1961	0.1308	0.0287	0.0282	0.0393	
		α	0.4256	0.0744	0.1489	0.0665	0.0638	0.0100	0.7990	0.4346	0.0654	0.1309	0.0093	0.0091	0.0044	
		ψ	0.4811	0.0329	0.0684	0.0507	0.0514	0.0036	0.9360	0.5198	0.0387	0.0804	0.0071	0.0074	0.0015	
P.D.	Exact	μ	1.9759	0.4759	0.3173	0.1868	0.2614	0.2600	1.7221	0.2221	0.1481	0.0283	0.0281	0.0501		
		α	0.3413	0.1587	0.3173	0.0604	0.0288	0.2420	0.4259	0.0741	0.1482	0.0092	0.0090	0.0056		
		ψ	0.4811	0.0756	0.1572	0.0504	0.0498	0.7090	0.5246	0.0435	0.0903	0.0072	0.0074	0.0019		
	Cond.	μ	1.9760	0.4760	0.3173	0.1872	0.1844	0.2616	0.2610	1.7221	0.2221	0.1481	0.0283	0.0281	0.0501	
		α	0.3413	0.1587	0.3174	0.0605	0.0594	0.0288	0.2450	0.4259	0.0741	0.1482	0.0092	0.0090	0.0056	
		ψ	0.4811	0.0755	0.1570	0.0505	0.0499	0.0083	0.7050	0.5246	0.0435	0.0903	0.0072	0.0074	0.0019	

^a True: True value of the corresponding parameter
^b Est.: Average of the estimate of the corresponding parameter across 1000 replications
^c Bias.abs: Absolute bias of the estimate
^d Bias.rel: Relative bias of the estimate
^e SE.emp: Empirical s.e. across 1000 replications
^f SE.avg: Average of the s.e. obtained from the model
^g MSE: Mean square error of the estimate, $MSE = Bias.abs^2 + SE.emp^2$.
^h Cover: Coverage probability of the estimate
ⁱ P.L.: Pooling likelihood functions
^j Exact: Maximizing the exact likelihood function
^k Cond.: Maximizing the conditional likelihood function
^l P.D.: Pooling data by connecting data directly

a pattern that the more categories, the more accurate the estimate. Within each table, by comparing the pooled likelihood methods (P.L.) and the pooling data methods (P.D.), the MSE values obtained from P.L. are in general smaller than those obtained from P.D., which indicated that the P.L. methods perform better than the P.D. methods. We further compare the exact-likelihood estimation method and the conditional-likelihood estimation method within each table. For the case of random $y_1 \sim N(\frac{\mu}{1-\alpha}, \frac{\psi}{1-\alpha^2})$, the pooled likelihood exact MLE are the best. For the other two initial values, the pooled likelihood conditional MLE are the best. In other word, the pooled likelihood conditional MLE is not sensitive to initial values.

5.2 Conclusions

Through the simulation, we have the following conclusions: (1) The pooled likelihood methods perform better than the pooling data methods. (2) For the case of random $y_1 \sim N(\frac{\mu}{1-\alpha}, \frac{\psi}{1-\alpha^2})$, the pooled likelihood exact MLE are the best. For the other two initial values, the pooled likelihood conditional MLE are the best. In other word, the pooled likelihood conditional MLE is not sensitive to initial values. (3) The more categories, the more accurate the estimate. (4) The longer the time series, the more accurate the estimate. (5) The more individual participated, the more accurate the estimate.

5.3 Discussion

In this article, the intra-individual variation for multiple subjects with categorical outcomes are examined. In reality, when multiple subjects are involved, methods developed for single subject intra-individual relationship may not fully work. Also, laws governing inter-individual relationship may not apply to intra-individual relationship, especially when outcomes are categorical, which are very common in social and behaviorial fields. The categorical or ordinal data are usually collected by the Likert table. There are few methods in literature dealing with the analysis of pooling multiple time series. Fewer for categorical data. This article fill the gap by investigating the performance of four estimation methods for pooling time series data focusing on categorical outcomes and to address related issues through an AR(1) model.

Appendix 1

```

##-----##
##      Data Generation Functions      ##
## Random  $y_1 \sim N(\mu/1-a, \text{sig}^2/1-a^2)$  ##
##-----##

ar1.ranI2.sim <- function(T, a, mu, sig, nc){
  ymean <- mu/(1-a)
  ysig  <- sig/(sqrt(1-a^2))
  et    <- rnorm(T, 0, sig)
  y     <- rep(1,T)
  y[1]  <- rnorm(1,ymean,ysig)
  yc    <- rep(1,T)

  ##-----##
  ## Categorical data
  ##-----##
  th    <- seq(-2, 2, length=nc-1)*ysig + ymean  ## thresholds
  categ <- seq(1, nc)                            ## categories

  ## For yc[1] when time=1
  if (y[1] <= th[1]) {yc[1] <- categ[1]}          ## s=1
  if (y[1] > th[nc-1]){yc[1] <- categ[nc]}       ## s=nc
  for (s in 2:(nc-1)){
    if ((th[s-1] < y[1])&(y[1] <= th[s]))
      {yc[1] <- categ[s]} }

  ## For yc[2:T]
  for (i in 2:T){
    temp <- mu+a*y[i-1]+et[i]
    if (temp <= th[1]) {yc[i] <- categ[1]}        ## s=1
    if (temp > th[nc-1]){yc[i] <- categ[nc]}     ## s=nc
    for (s in 2:(nc-1)){
      if ((th[s-1] < temp)&(temp <= th[s]))
        {yc[i] <- categ[s]} }
    }
  return(yc)}

##-----##
##      Model Estimation Functions      ##
##      Exact-likelihood Estimation     ##
##-----##

```

```

exactllike <- function(par, y){
  mu <- par[1]
  a <- par[2]
  psi <- par[3]
  N <- nrow(y)
  T <- ncol(y)
  Y1 <- as.vector(y[,1])
  Yt <- as.vector(y[,2:T])
  Yt1 <- as.vector(y[,1:(T-1)])
  sum <- (1-a^2)*t(Y1-mu/(1-a))%*(Y1-mu/(1-a))
    + t(Yt-mu-a*Yt1)%*(Yt-mu-a*Yt1)
  llik <- -.5*N*T*log(2*pi*psi) + .5*N*log(1-a^2) - sum/
    (2*psi)-llik}

##-----##
##      Model Estimation Functions      ##
## conditional-likelihood Estimation ##
##-----##
condllike <- function(par, y){
  mu <- par[1]
  a <- par[2]
  psi <- par[3]
  N <- nrow(y)
  T <- ncol(y)
  Yt <- as.vector(y[,2:T])
  Yt1 <- as.vector(y[,1:(T-1)])
  sum <- t(Yt-mu-a*Yt1)%*(Yt-mu-a*Yt1)
  llik <- -.5*N*(T-1)*log(2*pi*psi) - sum/(2*psi)
  -llik}

##-----##
## Core Code in Main Program of Estimation ##
##-----##
result <- nlm(exactllike, c(0,1/2,1/2), y, hessian=T)
result <- nlm(condllike, c(0,1/2,1/2), y, hessian=T)
se <- sqrt(diag(solve(result$hessian)))

y.conn <- as.vector(t(y))
result <- arima(y.conn, c(1,0,0), include.mean=T)
se <- sqrt(diag(result$var.coef))

```

Appendix 2

Table 2 Results for $c = 7$ categories

		True	$N = 50$					$T = 5$									
			Est.	Bias.abs	Bias.rel	SE.emp	SE.avg	MSE	Cover	Est.	Bias.abs	Bias.rel	SE.emp	SE.avg	MSE	Cover	
$y_1 = 0$																	
P.L.	Exact	μ	2	2.5456	0.5456	0.2728	0.2605	0.2427	0.3655	0.3930	2.1454	0.1454	0.0727	0.0379	0.0368	0.0226	0.0250
		α	0.5	0.3637	0.1363	0.2725	0.0624	0.0590	0.0225	0.3750	0.4636	0.0364	0.0727	0.0091	0.0088	0.0014	0.0220
		ψ	1.2053	0.9995	0.2058	0.1708	0.0957	0.0900	0.0515	0.3990	1.2230	0.0177	0.0147	0.0169	0.0173	0.0006	0.8370
	Cond.	μ	2	2.1549	0.1549	0.0775	0.3327	0.3135	0.1347	0.9160	2.1077	0.1077	0.0539	0.0387	0.0378	0.0131	0.1840
		α	0.5	0.4615	0.0385	0.0771	0.0792	0.0759	0.0078	0.9150	0.4731	0.0269	0.0539	0.0093	0.0090	0.0008	0.1440
		ψ	1.2053	1.2360	0.0307	0.0255	0.1190	0.1236	0.0151	0.9660	1.2478	0.0425	0.0353	0.0173	0.0178	0.0021	0.3150
P.D.	Exact	μ	2	2.6562	0.6562	0.3281	0.2461	0.2463	0.4912	0.2390	2.1465	0.1465	0.0733	0.0379	0.0371	0.0229	0.0250
		α	0.5	0.3360	0.1640	0.3280	0.0583	0.0594	0.0303	0.2070	0.4634	0.0366	0.0733	0.0091	0.0089	0.0014	0.0200
		ψ	1.2053	1.0425	0.1628	0.1350	0.1007	0.0933	0.0366	0.5590	1.2299	0.0246	0.0204	0.0170	0.0174	0.0009	0.7220
	Cond.	μ	2	2.6508	0.6508	0.3254	0.2472	0.2475	0.4846	0.2550	2.1464	0.1464	0.0732	0.0379	0.0371	0.0229	0.0250
		α	0.5	0.3373	0.1627	0.3254	0.0586	0.0597	0.0299	0.2190	0.4634	0.0366	0.0732	0.0091	0.0089	0.0014	0.0210
		ψ	1.2053	1.0467	0.1586	0.1316	0.1011	0.0938	0.0354	0.5690	1.2300	0.0247	0.0205	0.0170	0.0174	0.0009	0.7200
$y_1 \sim N(0, \psi)$																	
P.L.	Exact	μ	2	2.2378	0.2378	0.1189	0.2458	0.2459	0.1170	0.8650	2.1161	0.1161	0.0580	0.0370	0.0370	0.0148	0.1150
		α	0.5	0.4404	0.0596	0.1192	0.0586	0.0596	0.0070	0.8610	0.4709	0.0291	0.0582	0.0088	0.0088	0.0009	0.0850
		ψ	1.2053	1.1885	0.0168	0.0140	0.1005	0.1075	0.0104	0.9570	1.2408	0.0355	0.0295	0.0175	0.0176	0.0016	0.4610
	Cond.	μ	2	2.1587	0.1587	0.0794	0.2744	0.2715	0.1005	0.9170	2.1077	0.1077	0.0539	0.0374	0.0374	0.0130	0.1770
		α	0.5	0.4594	0.0406	0.0813	0.0639	0.0650	0.0057	0.9150	0.4730	0.0270	0.0541	0.0089	0.0089	0.0008	0.1480
		ψ	1.2053	1.2349	0.0296	0.0245	0.1183	0.1235	0.0149	0.9650	1.2463	0.0410	0.0340	0.0177	0.0178	0.0020	0.3690

P.D.	Exact	μ	2	2.5499	0.5499	0.2749	0.2449	0.2462	0.3624	0.4100	2.1461	0.1461	0.0730	0.0369	0.0372	0.0227	0.0170	
		α	0.5	0.3623	0.1377	0.2755	0.0577	0.0589	0.0223	0.3550	0.4634	0.0366	0.0732	0.0088	0.0089	0.0014	0.0070	
		ψ	1.2053	1.2892	0.0839	0.0696	0.1151	0.1153	0.0203	0.9200	1.2535	0.0482	0.0400	0.0177	0.0177	0.0026	0.2240	
		Cond.	μ	2	2.5487	0.5487	0.2744	0.2452	0.2466	0.3612	0.4100	2.1460	0.1460	0.0730	0.0369	0.0372	0.0227	0.0160
			α	0.5	0.3625	0.1375	0.2750	0.0577	0.0590	0.0222	0.3610	0.4634	0.0366	0.0732	0.0088	0.0089	0.0014	0.0070
		ψ	1.2053	1.2901	0.0848	0.0703	0.1154	0.1156	0.0205	0.9200	1.2535	0.0482	0.0400	0.0177	0.0177	0.0026	0.2240	
$y_1 \sim N\left(\frac{\mu}{1-\alpha}, \frac{\psi}{1-\alpha^2}\right)$																		
P.L.	Exact	μ	2	2.1288	0.1288	0.0644	0.2544	0.2439	0.0813	0.9220	2.1083	0.1083	0.0541	0.0373	0.0371	0.0131	0.1530	
		α	0.5	0.4673	0.0327	0.0654	0.0616	0.0591	0.0049	0.9310	0.4729	0.0271	0.0542	0.0090	0.0088	0.0008	0.1330	
		ψ	1.2053	1.2343	0.0290	0.0241	0.1095	0.1118	0.0128	0.9630	1.2471	0.0418	0.0347	0.0165	0.0176	0.0020	0.3280	
		Cond.	μ	2	2.1308	0.1308	0.0654	0.2714	0.2608	0.0908	0.9160	2.1083	0.1083	0.0542	0.0376	0.0373	0.0131	0.1630
			α	0.5	0.4669	0.0331	0.0662	0.0645	0.0623	0.0053	0.9200	0.4729	0.0271	0.0542	0.0090	0.0089	0.0008	0.1400
		ψ	1.2053	1.2293	0.0240	0.0199	0.1214	0.1230	0.0153	0.9600	1.2472	0.0419	0.0347	0.0166	0.0178	0.0020	0.3310	
P.D.	Exact	μ	2	2.5118	0.5118	0.2559	0.2527	0.2456	0.3258	0.4720	2.1462	0.1462	0.0731	0.0375	0.0372	0.0228	0.0230	
		α	0.5	0.3714	0.1286	0.2573	0.0603	0.0587	0.0202	0.4360	0.4634	0.0366	0.0732	0.0090	0.0089	0.0014	0.0160	
		ψ	1.2053	1.3635	0.1582	0.1312	0.1223	0.1220	0.0400	0.7790	1.2614	0.0561	0.0466	0.0169	0.0178	0.0034	0.0940	
		Cond.	μ	2	2.5116	0.5116	0.2558	0.2527	0.2459	0.3255	0.4780	2.1462	0.1462	0.0731	0.0375	0.0372	0.0228	0.0230
			α	0.5	0.3714	0.1286	0.2572	0.0602	0.0587	0.0202	0.4370	0.4634	0.0366	0.0732	0.0090	0.0089	0.0014	0.0170
		ψ	1.2053	1.3636	0.1583	0.1313	0.1225	0.1222	0.0401	0.7830	1.2614	0.0561	0.0466	0.0169	0.0178	0.0034	0.0940	

Note: With the same notations as in Table 1

Table 3 Results for $c = 9$ categories

		$N = 50$					$T = 5$											
True		Est.	Bias.abs	Bias.rel	SE.emp	SE.avg	MSE	Cover	Est.	Bias.abs	Bias.rel	SE.emp	SE.avg	MSE	Cover			
$y_1 = 0$	P.L.	Exact	μ	2.5	3.1544	0.6544	0.2618	0.3066	0.3038	0.5222	0.4300	2.6259	0.1259	0.0503	0.0461	0.0180	0.2160	
			α	0.5	0.3684	0.1316	0.2633	0.0597	0.0588	0.0209	0.3940	0.3940	0.4747	0.0253	0.0507	0.0087	0.0007	0.1680
			ψ	2.2750	1.8701	0.4049	0.1780	0.1839	0.1684	0.1978	0.3500	2.2738	0.0012	0.0005	0.0305	0.0322	0.0009	0.9580
		Cond.	μ	2.5	2.6606	0.1606	0.0642	0.3922	0.0756	0.0068	0.9290	2.5777	0.0157	0.0314	0.0089	0.0090	0.0003	0.6250
			α	0.5	0.4668	0.0332	0.0664	0.0758	0.2312	0.0533	0.3093	2.3199	0.0449	0.0197	0.0311	0.0331	0.0030	0.5870
	P.D.	Exact	ψ	2.2750	2.3115	0.0365	0.0160	0.2280	0.3107	0.7016	2.6271	0.1271	0.0508	0.0463	0.0465	0.0184	0.2160	
			μ	2.5	3.2921	0.7921	0.3168	0.2903	0.0284	0.2070	0.4744	0.0256	0.0513	0.0087	0.0088	0.0007	0.1720	
			α	0.5	0.3409	0.1591	0.3183	0.0557	0.1747	0.1414	0.5130	2.2874	0.0124	0.0054	0.0307	0.0324	0.0011	0.9510
		Cond.	ψ	2.2750	1.9523	0.3227	0.1418	0.1930	0.3107	0.7016	2.6271	0.1271	0.0508	0.0463	0.0465	0.0183	0.2190	
			μ	2.5	3.2853	0.7853	0.3141	0.2915	0.0596	0.0280	0.2170	0.4744	0.0256	0.0512	0.0087	0.0088	0.0007	0.1740
	α	0.5	0.3422	0.1578	0.3155	0.0559	0.1757	0.1367	0.5280	2.2876	0.0126	0.0055	0.0307	0.0324	0.0011	0.9500		
	ψ	2.2750	1.9601	0.3149	0.1384	0.1938												
$y_1 \sim N(0, \psi)$	P.L.	Exact	μ	2.5	2.7244	0.2244	0.0898	0.3084	0.3059	0.1455	0.9160	2.5894	0.0894	0.0357	0.0474	0.0102	0.5060	
			α	0.5	0.4561	0.0439	0.0877	0.0596	0.0589	0.0055	0.9000	0.4822	0.0178	0.0356	0.0090	0.0088	0.0004	0.4690
			ψ	2.2750	2.2075	0.0675	0.0297	0.1886	0.1998	0.0401	0.9330	2.3106	0.0356	0.0157	0.0302	0.0327	0.0022	0.8350
		Cond.	μ	2.5	2.6261	0.1261	0.0504	0.3371	0.3399	0.1295	0.9470	2.5785	0.0785	0.0314	0.0481	0.0469	0.0085	0.6190
			α	0.5	0.4757	0.0243	0.0486	0.0642	0.0644	0.0047	0.9470	0.4843	0.0157	0.0313	0.0091	0.0089	0.0003	0.5810
	P.D.	Exact	ψ	2.2750	2.2932	0.0182	0.0080	0.2134	0.2294	0.0459	0.9640	2.3210	0.0460	0.0202	0.0306	0.0332	0.0030	0.7410
			μ	2.5	3.1381	0.6381	0.2552	0.3117	0.3095	0.5044	0.4760	2.6278	0.1278	0.0511	0.0477	0.0466	0.0186	0.2320
			α	0.5	0.3735	0.1265	0.2530	0.0594	0.0586	0.0195	0.4470	0.4745	0.0255	0.0510	0.0091	0.0088	0.0007	0.1860
		Cond.	ψ	2.2750	2.4139	0.1389	0.0611	0.2081	0.2160	0.0626	0.9340	2.3357	0.0607	0.0267	0.0307	0.0330	0.0046	0.5610
			μ	2.5	3.1368	0.6368	0.2547	0.3122	0.3099	0.5031	0.4750	2.6277	0.1277	0.0511	0.0477	0.0466	0.0186	0.2310
	α	0.5	0.3737	0.1263	0.2526	0.0594	0.0587	0.0195	0.4480	0.4745	0.0255	0.0510	0.0091	0.0088	0.0007	0.1870		
	ψ	2.2750	2.4151	0.1401	0.0616	0.2081	0.2165	0.0629	0.9370	2.3358	0.0608	0.0267	0.0307	0.0330	0.0046	0.5600		

		$y_1 \sim N\left(\frac{\mu}{1-\alpha}, \frac{\psi}{1-\alpha^2}\right)$															
P.L.	Exact	μ	2.5	2.6150	0.1150	0.0460	0.3038	0.3045	0.1055	0.9470	2.5760	0.0760	0.0304	0.0461	0.0464	0.0079	0.6230
		α	0.5	0.4773	0.0227	0.0454	0.0580	0.0586	0.0039	0.9520	0.4848	0.0152	0.0305	0.0087	0.0088	0.0003	0.5950
		ψ	2.2750	2.3029	0.0279	0.0123	0.1964	0.2088	0.0394	0.9560	2.3183	0.0433	0.0191	0.0301	0.0328	0.0028	0.7660
	Cond.	μ	2.5	2.6147	0.1147	0.0459	0.3191	0.3280	0.1150	0.9460	2.5759	0.0759	0.0304	0.0466	0.0468	0.0079	0.6340
		α	0.5	0.4773	0.0227	0.0454	0.0598	0.0620	0.0041	0.9530	0.4848	0.0152	0.0304	0.0088	0.0088	0.0003	0.5970
		ψ	2.2750	2.2954	0.0204	0.0089	0.2152	0.2296	0.0467	0.9570	2.3184	0.0434	0.0191	0.0304	0.0331	0.0028	0.7720
	P.D.	μ	2.5	3.0991	0.5991	0.2396	0.3099	0.3092	0.4550	0.5180	2.6247	0.1247	0.0499	0.0461	0.0466	0.0177	0.2370
		α	0.5	0.3804	0.1196	0.2391	0.0582	0.0584	0.0177	0.4780	0.4750	0.0250	0.0499	0.0087	0.0088	0.0007	0.1910
		ψ	2.2750	2.5538	0.2788	0.1226	0.2207	0.2285	0.1265	0.8320	2.3467	0.0717	0.0315	0.0304	0.0332	0.0061	0.4010
	Cond.	μ	2.5	3.0995	0.5995	0.2398	0.3093	0.3096	0.4551	0.5190	2.6247	0.1247	0.0499	0.0461	0.0466	0.0177	0.2370
		α	0.5	0.3804	0.1196	0.2391	0.0582	0.0585	0.0177	0.4780	0.4750	0.0250	0.0499	0.0087	0.0088	0.0007	0.1910
		ψ	2.2750	2.5540	0.2790	0.1226	0.2208	0.2289	0.1266	0.8300	2.3467	0.0717	0.0315	0.0304	0.0332	0.0061	0.3970

Note: With the same notations as in Table 1

References

- Cattell, R. B. (1952). The three basic factor-analytic research designs-their interrelations and derivatives. *Psychological Bulletin*, *49*, 499–520.
- Cattell, R. B., Cattell, A. K. S., & Rhymer, R. M. (1947). P-technique demonstrated in determining psychophysical source traits in a normal individual. *Psychometrika*, *12*, 267–288.
- Cattell, R. B., & Scheier, I. H. (1961). *The meaning and measurement of neuroticism and anxiety*. New York: Ronald Press.
- Daly, D. L., Bath, K. E., & Nesselroade, J. R. (1974). On the confounding of interand intraindividual variability in examining change patterns. *Journal of Clinical Psychology*, *30*, 33–36.
- Molenaar, P. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, *50*, 181–202.
- Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology – this time forever. *Measurement: Interdisciplinary Research and Perspectives*, *2*, 201–218.
- Molenaar, P. C. M., Huizenga, H. M., & Nesselroade, J. R. (2003). The relationship between the structure of interindividual and intraindividual variability: A theoretical and empirical vindication of developmental systems theory. In U. M. Staudinger & U. Lindenberger (Eds.), *Understanding human development: Dialogues with lifespan psychology* (pp. 339–360). Norwell: Kluwer.
- Nesselroade, J. R., & Molenaar, P. (2003). Quantitative models for developmental processes. In J. Valsiner & K. Connolly (Eds.), *Handbook of developmental psychology* (pp. 622–639). London: Sage.
- Nesselroade, J. R., & Molenaar, P. C. M. (1999). Pooling lagged covariance structures based on short, multivariate time-series for dynamic factor analysis. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 224–250). Newbury Park: Sage.
- Nesselroade, J. R., & Ram, N. (2004). Studying intraindividual variability: What we have learned that will help us understand lives in context. *Research in Human Development*, *1*, 9–29.

An Investigation of Prior Specification on Parameter Recovery for Latent Dirichlet Allocation of Constructed-Response Items



Jordan M. Wheeler , Jiawei Xiong , Constanza Mardones-Segovia , Hye-Jeong Choi, and Allan S. Cohen 

Abstract Latent Dirichlet Allocation (LDA) is a probabilistic model to analyze textual data. It was originally developed for corpora containing large amount of textual data, such as large sets of journal abstracts, blogs, and newspaper articles. Recently, LDA has been applied in psychological and educational measurement fields to analyze examinees' responses to open-ended items on assessments. The amount of textual data found in educational measurement scenarios, however, is notably less than the amount of data originally used for LDA. The observed data, therefore, may not be enough to accurately recover the parameters. Thus, it is important to explore how various priors influence the parameter recovery of the LDA model. In this study, we investigated the effects of prior hyperparameters parameter on recovery through a simulation using various conditions that are common in educational assessment settings. Specifically, five sets of priors ranging from highly informative to noninformative were used. For each set of priors, four factors were manipulated and all factors were crossed for a total of 108 conditions. The four factors used in this study were: number of unique words (3 levels: 250, 500, and 750 words), average response length (3 levels: 5, 25, and 50 words per document), number of documents (3 levels: 100, 250, and 500 documents), and number of topics (3 levels: 3, 4, and 5 topics). The results of the simulation showed that the prior specification of the LDA model influenced the parameter recovery rates.

Keywords Educational topic models · Model recovery · Simulation study

J. M. Wheeler (✉) · J. Xiong · C. Mardones-Segovia, · A. S. Cohen
The University of Georgia, Athens, GA, USA

e-mail: jmwheeler@uga.edu; Jiawei.Xiong@uga.edu; cam04214@uga.edu; acohen@uga.edu

H.-J. Choi

The University of Georgia, Athens, GA, USA

The Human Resources Research Organization, Louisville, KY, USA

e-mail: hjchoi1@uga.edu

1 Introduction

Latent Dirichlet Allocation (LDA; Blei et al., 2003) is a type of topic model that uses a probabilistic framework to estimate a predetermined number of latent topics within a collection of documents. The estimated latent topics are clusters of words that characterize common sets of words that are seen throughout the collection of documents. Each document within the collection is given a set of proportions that expresses the usage of each latent topic.

LDA, along with other topic models, was originally developed for large corpora as a way to easily index and characterize each document. More recently, LDA has been used in educational measurement as a method to analyze responses to constructed-response (CR) items. The results of this method provide researchers with a set of topics that are used across the responses for a particular CR item. Additionally, this method provides researchers with a set of proportions that show relatively usage of the topics for each response, which can be used to classify or cluster subsets of responses.

In educational measurement, LDA has been used in a wide array of studies. Choi et al. (2019) and Xiong et al. (2019) used LDA to analyze a set of middle-grade CR items and investigated the relationship between the topics and scores. Kim et al. (2017) showed that the results from the LDA analysis on CR item responses provided additional useful information about examinees beyond the scores. Duong et al. (2019) showed that LDA can be used to capture the effects of instructional interventions. LDA has also been used to investigate rater accuracy to determine potential reasons why some responses were more difficult to score accurately than others (Wheeler et al., 2022).

One issue is that the interpretation of the results from LDA in the above studies depended on selecting an appropriate number of latent topics to estimate, which is determined by the researchers and not the LDA model. In these studies, researchers employ various model selection indices and methods that help determine the most appropriate number of topics for a given set of responses. The number of topics selected for an LDA model depend on both the model selection method and the prior specification. Mardones et al. (2021) conducted a simulation study that evaluated the effectiveness of various model selection methods used for LDA under realistic educational conditions. The researchers noted that there seemed to be an effect of prior specification on the performance of the various model selection methods, however, it was not the main focus of their study. Furthermore, there has been no study that specifically looked into the effect of prior specification for the LDA model with educational data. This study, therefore, conducts a simulation study to investigate the effects of prior specification on parameter recovery of the LDA model.

1.1 Purpose of Study

The purpose of this study was to evaluate the effect of prior specifications on parameter recovery of the LDA model. Specifically, this study investigated two questions: (1) Does prior specification for the LDA model significantly impact the recovery of model parameters? and (2) what types of prior hyperparameters perform best under realistic measurement settings? The paper is structured as follows. First, the LDA model is presented with emphasis on the influence the prior hyperparameters have on the estimated parameters. Next, a simulation study is presented to investigate the effects of prior hyperparameters on parameter recovery under common constructed-response scenarios. The results of the simulation study are presented and an analysis of variance is conducted to evaluate which manipulated factors had a significant influence on the parameter recovery. Finally, the study is concluded by discussing the practical implications of the results of the simulation study.

2 Latent Dirichlet Allocation

The LDA model estimates three main parameters from a corpus of documents: a set of word probabilities (i.e., topics) β , a set of topic proportions for individual documents θ , and a set of topic assignments z . Each topic is a set of probabilities over the vocabulary V , which is the set of unique words found across all documents. The set of probabilities that constitute a topic express the probabilities of each word from the vocabulary appearing under the given topic. Each document is given a set of topic proportions which is a vector of proportions over the topics. The topic proportions express the proportion of each topic being used in the given document. Additionally, each document is given a set of topic assignments that represent the topic membership of each word that appears in the document.

The LDA model is a hierarchical mixture model where topics are considered corpus-wide parameters, topic proportions are considered document-wide parameters, and topic assignments are considered word-wide parameters. The mixture component of LDA is that topics are a mixture of words from the vocabulary and topic proportions are a mixture of topics. Additionally, LDA assumes the following generative process: (1) assume that there are K topics, (2) assume each document is generated by first generating its topic proportion, (3) assume a topic is assigned to each word in the document which is determined by its topic proportion, (4) assume each word in the document is generated given the topic assignment and topic distribution.

2.1 Mathematical Formulation

Suppose that there is a corpus that contains $d = 1, 2, \dots, D$ documents, each document contains N_d words, and there are V unique words across all documents. The LDA model assumes a priori that there are K corpus-wide topics. The joint distribution for the observed word variables ($\mathbf{w}_{1:D}$), the latent topic assignment variables ($\mathbf{z}_{1:D}$), the topics ($\boldsymbol{\beta}_{1:K}$), and the topic proportions ($\boldsymbol{\theta}_{1:D}$) is given by

$$p(\mathbf{w}_{1:D}, \mathbf{z}_{1:D}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\theta}_{1:D} | \boldsymbol{\eta}, \boldsymbol{\nu}), \quad (1)$$

where $\boldsymbol{\eta}$ and $\boldsymbol{\nu}$ are the prior hyperparameters. The joint distribution for LDA can subsequently be factorized into the conditional distributions (i.e., likelihood of the data) and priors, which is shown in

$$\begin{aligned} p(\mathbf{w}_{1:D}, \mathbf{z}_{1:D}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\theta}_{1:D} | \boldsymbol{\eta}, \boldsymbol{\nu}) = \\ p(\mathbf{w}_{1:D} | \mathbf{z}_{1:D}, \boldsymbol{\beta}_{1:K}) p(\mathbf{z}_{1:D} | \boldsymbol{\theta}_{1:D}) p(\boldsymbol{\beta}_{1:K} | \boldsymbol{\nu}) p(\boldsymbol{\theta}_{1:D} | \boldsymbol{\eta}), \end{aligned} \quad (2)$$

where $p(\mathbf{w}_{1:D} | \mathbf{z}_{1:D}, \boldsymbol{\beta}_{1:K})$ is the conditional distribution of the observed words given the topic assignments and the topics and is assumed to follow a multinomial distribution; $p(\mathbf{z}_{1:D} | \boldsymbol{\theta}_{1:D})$ is the conditional distribution of topic assignments given the topic proportions and is assumed to follow a multinomial distribution; $p(\boldsymbol{\beta}_{1:K} | \boldsymbol{\nu})$ is the prior distribution for the topics where $\boldsymbol{\nu}$ is the prior hyperparameter which controls the density of the word probabilities and is assumed to follow a Dirichlet distribution; and $p(\boldsymbol{\theta}_{1:D} | \boldsymbol{\eta})$ is the prior distribution for the topic proportions where $\boldsymbol{\eta}$ is the prior hyperparameter which controls the density of topic proportions and is assumed to follow a Dirichlet distribution (Blei et al., 2003; Ponweiser, 2012).

The prior hyperparameters, $\boldsymbol{\nu}$ and $\boldsymbol{\eta}$, are V -dimensional and K -dimensional vector space parameters, respectively, such that $\boldsymbol{\nu} = [\nu_1, \nu_2, \dots, \nu_V]$ and $\boldsymbol{\eta} = [\eta_1, \eta_2, \dots, \eta_K]$. The values within the $\boldsymbol{\nu}$ and $\boldsymbol{\eta}$ vectors influence the estimates for the topic and topic proportion parameters, respectively. When the hyperparameter for the topics are set to small values, such as $\nu_1, \nu_2, \dots, \nu_V < 1$, then this causes the estimated topics to have large probabilities for a small subset of words and small probabilities for the other words. Similarly, when the hyperparameter for the topics are set to small values, such as $\eta_1, \eta_2, \dots, \eta_K < 1$, then this causes the estimated topic proportions to estimate high proportions for few topics and small proportions for the other topics. When the hyperparameter for the topics are set to large values, such as $\nu_1, \nu_2, \dots, \nu_V > 1$, then this causes the estimated topics to have evenly spread probabilities for majority of the words. Similarly, when the hyperparameter for the topics are set to large values, such as $\eta_1, \eta_2, \dots, \eta_K > 1$, then this causes the estimated topic portions to have evenly spread proportions for majority of all topics. Specifying the hyperparameters for the two priors in LDA may impact the estimated parameters, especially when datasets are smaller (i.e., fewer documents, fewer unique words, fewer number of topics, and smaller average

document length); therefore, it is important to consider how the prior specification will influence parameter estimation.

Since the joint distribution is intractable due to its dimensionality, LDA infers the latent parameters $(z_{1:D}, \beta_{1:K}, \theta_{1:D})$ through the conditional distribution of these parameters given the observed words: $p(z_{1:D}, \beta_{1:K}, \theta_{1:D} | w_{1:D})$. This distribution is referred to as the posterior distribution, which is proportional to the joint distribution, given by

$$p(z_{1:D}, \beta_{1:K}, \theta_{1:D} | w_{1:D}) \propto p(w_{1:D}, z_{1:D}, \beta_{1:K}, \theta_{1:D}), \tag{3}$$

where $p(w_{1:D}, z_{1:D}, \beta_{1:K}, \theta_{1:D})$ is the joint distribution shown in Eq. (2). Given the dimensionality of the corpus, the posterior distribution can be factorized as

$$p(z_{1:D}, \beta_{1:K}, \theta_{1:D} | w_{1:D}) \propto \prod_{k=1}^K P(\beta_k | v) \prod_{d=1}^D \left(P(\theta_d | \eta) \prod_{n=1}^{N_d} P(w_{d,n}, z_{d,n} | \theta_d, \beta) \right). \tag{4}$$

The posterior distribution of LDA can be estimated through various techniques. The two primary methods for estimating the posterior is through Gibbs sampling and variational inference (Blei et al., 2003).

2.2 Generative Model

In addition to the joint distribution for the LDA Model, there is also an assumed generative process. The LDA model assumes that a document is generate through the following process with D documents and K topics (Blei et al. 2003; Blei, Ng, & Jordan 2002):

For each document $d = 1, 2, \dots, D$,

1. Choose $N_d \sim \text{Poisson}(\lambda)$
2. Choose $\theta_d \sim \text{Dirichlet}(\eta)$
3. For each word $n \in 1, 2, \dots, N_d$:
 - i. Choose $z_{d,n} \sim \text{Multinomial}(\theta_d)$
 - ii. Choose $w_{d,n} \sim \text{Multinomial}(\beta_k | z_{d,n})$

where N_d is the number of words in document d ; λ is the average document length; θ_d is the topic proportions of document d , $z_{d,n}$ is the topic assignment for n th word in document d , $w_{d,n}$ is the generated observed text for the n th word in document d ; and β_k are the word probabilities for the k th topic. The estimation algorithms use this generative process to infer the latent parameters $\theta_{1:D}$, $\beta_{1:K}$, and $z_{1:D}$.

2.3 *Requirements for Analysis*

There are a few requirements to obtain interpretable features from latent Dirichlet allocation. First, to obtain features that are representative of the documents, word that are high frequency and have low meaning should be removed. These words are often referred to as stop words and consist of words such as: *the, and, a, it, was, is*. Second, to extract meaningful features using latent Dirichlet allocation, the same words from different tenses must be converted to the same tense. This process is referred to as stemming and converts words such as *running, ran, runs, run* into the same word/tense *run*. These two data preparation requirements increase the interpretability and quality of the features extracted (Schofield et al., 2017). Beyond removing stop words and stemming, there are no requirements for extracting features using latent Dirichlet allocation except for the words and the documents that they belong to. The word data is converted into vectors where each word is associated with a number and each position in the vector is associated with a document.

3 *Simulation Design*

The impact of prior specification on parameter recovery for the LDA model was investigated using a simulation study using 30 replications. For each condition of the simulation study, an LDA model was estimated using a collapsed Gibbs sampling algorithm (Griffiths & Steyvers 2004, Hoffman, Bach, & Blei 2010). The topic parameter estimates and the topic proportion parameter estimates estimated were compared to the generating parameters to determine how well the model recovered the parameters. The following subsections explain the simulation design and conditions, the data generation mechanism, the software used to estimate the model, and the evaluation criteria.

3.1 *Simulation Design and Conditions*

There were a total of five factors manipulated in the study: number of topics, vocabulary length, number of documents, average document length, and prior hyperparameters. Table 1 shows the levels for each factor manipulated. The levels for the vocabulary length, the average essay length, the number of topics, and the number of documents were determined from previous simulation studies involving LDA (Mardones et al. 2021; Wheeler et al. 2021). The average essay length condition represents constructed-response items that require shorter responses and the number of documents condition represents school and district size samples. The prior hyperparameters that were used in the simulation study were chosen based

Table 1 The manipulated factors for the simulation study

Factor	Number of levels	Levels
Vocabulary length	3	250, 500, 750
Average document length	4	5, 25, 50
Number of topics	3	3, 4, 5
Number of document	3	100, 250, 500
Prior hyperparameters	5	$\nu = 5$ and $\eta = 5$, $\nu = 1$ and $\eta = 1$, $\nu = 0.5$ and $\eta = 0.5$, $\nu = 1/K$ and $\eta = 1/V$, $\nu = 0.1$ and $\eta = 0.1$

on values used from previous applications of LDA to educational data (e.g., Kwak 2019; Xiong et al. 2019; Xiong et al. 2021). The computational cost of estimating the LDA model is expensive due to the total number of parameters in the model, therefore, only 30 replications were used in this study. For each replication, all conditions were crossed and the topic and topic proportion parameter estimates were compared to the generating parameters to evaluate the degree of recovery.

3.2 Data Generation

To generate the data for this simulation study, we followed the generative model shown in the previous section. One thing that Choi et al. (2020) noted is that the topics estimated from responses to CR items often overlap each other. That is, when students are prompted with a CR item, they are constrained to what they can write about, which constrains the estimated topics and can often cause them to overlap. To emulate this in our simulation study, we used real data from a middle-grade English Language Arts CR item to derive the generating parameters for the topics. Specifically, we used the real data to estimate the LDA model and used the estimated topics, which overlapped, as the generating Dirichlet distribution for our simulation study.

3.3 Model Estimation

The joint posterior distribution specified in Eq. (2) is intractable, but it can be estimated through various techniques, such as the collapsed Gibbs sampling algorithm. The collapsed Gibbs sampling method was used to estimate the LDA using the simulated data. We used the R programming language to first generate the data and then used the *topicmodels* (Hornik & Grün, 2011) and *lda* (Chang, 2015) packages to implement the collapsed Gibbs sampling algorithm for LDA to estimate the parameters from the generated data.

3.4 Evaluation of Parameter Recovery

Cosine similarity and root mean squared error (RMSE) are used to measure the performance of parameter recovery of the word probabilities (i.e., topics). The cosine similarity measure is given in Eq. (2),

$$\cos(\hat{\beta}, \beta) = \frac{\sum_{v=1}^V \hat{\beta}_v \cdot \beta_v}{\sqrt{\sum_{v=1}^V \hat{\beta}_v^2} \cdot \sqrt{\sum_{v=1}^V \beta_v^2}}, \quad (5)$$

where $\hat{\beta}$ is the estimated word probability; β is the known generated word probability; and V is the length of the vocabulary. Cosine similarity is a type of correlation measure and is often used for large dimensional vectors. Within a topic modeling context, cosine similarity measures the similarity between two topics (Singhal, 2001). In this study, the cosine similarity is measuring how similar the estimated topic is to the known generated topic. A cosine similarity near 1 indicates that the estimated and known topics are similar, whereas a cosine similarity near 0 indicates that the estimated and known topics are not similar. Previous LDA studies use a cosine similarity value of 0.8 or higher to indicate parameter recovery (Wheeler et al., 2021).

Additionally, RMSE is used to evaluate the recovery of word probabilities. The RMSE measure for word probability recovery is given in Eq. (3),

$$RMSE = \sqrt{\frac{\sum_{v=1}^V (\hat{\beta}_v - \beta_v)^2}{V}}, \quad (6)$$

where $\hat{\beta}$ is the estimated word probability; β is the known generated word probability; and V is the length of the vocabulary. RMSE is a measure that indicates a relative error. An RMSE of 0 mean that there was perfect recovery for the topics, therefore, a smaller RMSE indicates better recovery. In this study, we also used RMSE to evaluate the recovery of topic proportions. The RMSE measure for topic proportion recovery is given in Eq. (2),

$$RMSE = \sqrt{\frac{\sum_{k=1}^K (\hat{\theta}_k - \theta_k)^2}{K}}, \quad (7)$$

where $\hat{\theta}_k$ is the estimated topic proportion for topic k ; θ_k is the known topic proportion for topic k ; and K is the number of topics. Similar to RMSE of word probabilities, a smaller RMSE indicates better recovery. That is, a smaller RMSE

between the estimated topic proportions and the generated topic proportions indicates that the LDA model successfully recovered the topic proportion parameters.

Additionally, we conducted an analysis of variance for word probabilities and topic proportions. For the analysis of variance for word probabilities, we used the RMSE to determine which simulation factors had significant influence on parameter recovery. Similarly, for the analysis of variance of topic proportions, we used the RMSE values to determine which simulation factors had significant influence on parameter recovery.

4 Results

For each condition of the simulation study, 30 replications were conducted. The average RMSE of word probabilities (topic distributions) and topic proportions were calculated across the 30 replications for each condition. An analysis of variance for the manipulated factors according to the average RMSE for both the word probabilities and topic proportions was conducted. Table 2 shows the result of the analysis of variance for the average RMSE of word probabilities. Additionally, an effect size (partial η^2) was calculated for each factor.

From Table 2, it can be seen that all factors that were manipulated for this simulation study had a significant effect on the RMSE of word probabilities. Furthermore, it can be seen that the vocabulary length and prior hyperparameter conditions had the largest effect sizes (partial $\eta^2 = 0.77$ and partial $\eta^2 = 0.62$, respectively). This result suggests that researchers should consider the number of unique words within the corpus when selecting an appropriate prior hyperparameter for the word probabilities.

Table 2 Analysis of variance for simulation factors according to the average RMSE of word probabilities (i.e., topics)

Sources of variation	Df	Sum sq ($\times 10^{-2}$)	Mean sq ($\times 10^{-2}$)	F value	P value	Partial η^2
Vocabulary length	2	0.0800	0.0400	649.55	<0.001	0.77
Prior hyperparameters	4	0.0389	0.0097	158.01	<0.001	0.62
Average document length	2	0.0142	0.0071	115.11	<0.001	0.37
Number of topics	2	0.0138	0.0069	112.16	<0.001	0.36
Number of documents	2	0.0083	0.0042	67.62	<0.001	0.26
Residuals	392	0.3505	0.0009			

Figure 1 shows two plots for the average cosine similarity between the estimated and true word probabilities. The left-hand plot shows the results when there are 3 topics, a vocabulary length of 250, and an average document length of 5. The right-hand plot shows the results when there are 3 topics, a vocabulary length of 250, and an average document length of 50. From the two plots, it can be seen that a prior hyperparameter for word probabilities around 1 ($\nu = 0.5$ or $\nu = 1$) performed better regardless of the number of documents or average document length in the corpus.

Table 3 shows the result of the analysis of variance for the average RMSE of topic proportions. Additionally, an effect size (partial η^2) was calculated for each factor. From Table 3, it can be seen that all factors that were manipulated for this simulation study had a significant effect on the RMSE of topic proportions. Furthermore, it can be seen that the prior hyperparameter conditions had the largest effect size (partial $\eta^2 = 0.67$), and that the vocabulary length and number of document conditions had relatively small effect sizes (partial $\eta^2 = 0.02$ and partial $\eta^2 = 0.02$, respectively).

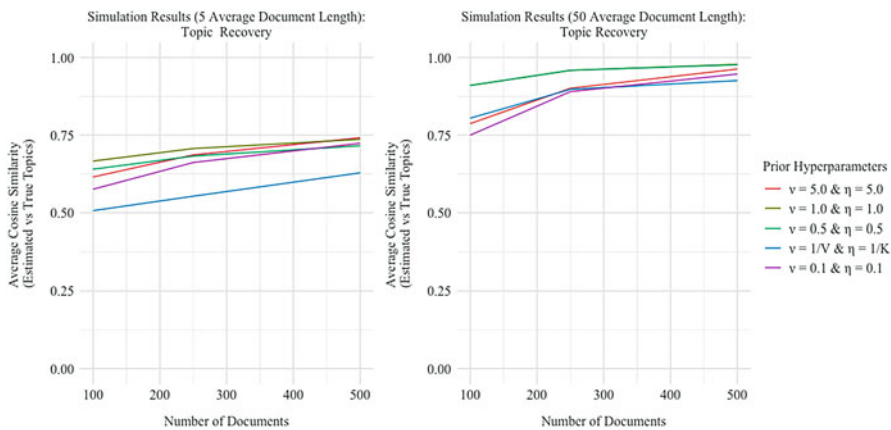


Fig. 1 Side by side plots for the average cosine similarity between estimated and true word probabilities (i.e., topics) under various simulation conditions. *Note.* Simulation condition for left hand plot: number of topics = 3, vocabulary length =250, and average document length = 5; Simulation condition for right hand plot: number of topics = 3, vocabulary length =250, and average document length = 50

Table 3 Analysis of variance for simulation factors according to the RMSE of topic proportions

Sources of variation	Df	Sum sq	Mean sq	F value	P Value	Partial η^2
Prior hyperparameters	4	0.7092	0.1773	198.31	<0.001	0.67
Average document length	2	0.2197	0.1098	122.82	<0.001	0.39
Number of topics	2	0.0791	0.0395	44.21	<0.001	0.18
Vocabulary length	2	0.0074	0.0037	4.16	0.016	0.02
Number of documents	2	0.0060	0.0030	3.36	0.036	0.02
Residuals	392	0.3505	0.0009			

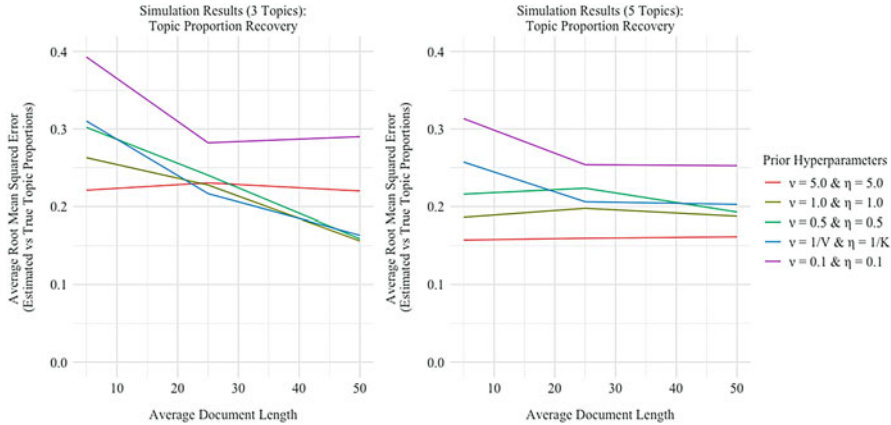


Fig. 2 Side by side plots for the average RMSE between estimated and true topic proportions under various simulation conditions. *Note.* Simulation condition for left hand plot: number of topics = 3, vocabulary length =250, and average document length = 5; Simulation condition for right hand plot: number of topics = 3, vocabulary length =250, and average document length = 50

The results from Table 3 suggest that the recovery of topic proportions does not heavily rely on the vocabulary length nor number of documents and that the prior hyperparameters have a large influence on parameter recovery. Furthermore, this result suggests that researchers need to carefully consider an appropriate prior hyperparameter for topic proportions if they want to accurately recover the true parameters.

Figure 2 shows two plots for the average RMSE between the estimated and true topic proportions. The left-hand plot shows the results when there are 3 topics, a vocabulary length of 250, and 100 documents. The right-hand plot shows the results when there are 5 topics, a vocabulary length of 250, and 100 documents. From the two plots, it can be seen that a larger prior hyperparameter for topic proportions ($\eta = 5$) when the average document length is small (~ 5) performed better regardless of the number of topics. Additionally, when the average document length increases for smaller number of topics, a prior hyperparameter for topic proportions near 1 (i.e, $\eta = 0.5$ or $\eta = 1$) resulted in better parameter recovery.

5 Discussion

This study investigated the effects of prior specification on parameter recovery for the LDA model under conditions that are often seen in educational measurement settings. The results of the simulation study show that the prior specification does significantly influence parameter estimation. Additionally, the other factors manipulated in this study also had significant effects on the parameter estimates.

This suggests that when fitting an LDA model to measurement data, the researchers must consider not only the prior hyperparameters but also the characteristics of the corpus, such as vocabulary length, number of documents, average document length, and number of topics.

Based on the analysis of variance on simulation conditions for the RMSE of word probabilities (i.e., topic distributions), the vocabulary length and prior hyperparameters had the biggest effect sizes. This suggests that when specifying the prior hyperparameter for topics, one should consider the number of unique words in the corpus being analyzed. Furthermore, Fig. 1 shows the average cosine similarity for topic recovery and suggests that a prior hyperparameter near 1 ($\nu_{1:V} = 1$) performed the best under most conditions. Based on the analysis of variance on simulation conditions for the RMSE of topic proportions, prior hyperparameters and average document length had the biggest effect sizes. This suggests that when specifying the prior hyperparameter for topics, one should consider the average length of the documents in the corpus being analyzed. Furthermore, the average RMSE for topic proportion recovery suggest that as the average length of a document increases, then a prior hyperparameter for topic proportions near 1 ($\eta_{1:K} \sim 1$) performed better. Additionally, the results also suggest that a larger prior hyperparameter for topic proportions ($\eta_{1:K} > 1$) performed best when there were more topics.

References

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2002). Latent dirichlet allocation. In *Advances in Neural Information Processing Systems* (pp. 601–608).
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Chang, J. (2015). *Lda: Collapsed Gibbs sampling methods for topic models* [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=lda> (R package version 1.4.2)
- Choi, H.-J., Kim, S., Cohen, A. S., Templin, J., & Copur-Gencturk, Y. (2020). Integrating statistical topic models and a diagnostic classification model for analyzing items in a mixed format assessment. *Frontiers in Psychology*, 11, 3997.
- Choi, H.-J., Kwak, M., Kim, S., Xiong, J., Cohen, A. S., & Bottge, B. A. (2019). An application of a topic model to two educational assessments. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative Psychology: 83rd Annual Meeting of the Psychometric Society* (vol. 265, pp. 449–459). New York: Springer.
- Duong, E., Mellom, P., & Hixon, R. (2019). Using topic modeling to analyze the effects of instructional conversation on 3rd grade students' writing. In *Paper Presented at the Annual Meeting of the American Association for Applied Linguistics, Atlanta*.
- Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1), 5228–5235.
- Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems* (pp. 856–864).
- Hornik, K., & Grün, B. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13), 1–30.

- Kim, S., Kwak, M., Cardozo-Gaibisso, L., Buxton, C., & Cohen, A. S. (2017). Statistical and qualitative analyses of students' answers to a constructed response test of science inquiry knowledge. *Journal of Writing Analytics, 1*(1), 82–102.
- Kwak, M. (2019). *Parameter Recovery in Latent Dirichlet Allocation (lda): Potential Utility of lda in Formative Constructed Response Assessment*. (Unpublished doctoral dissertation). University of Georgia.
- Mardones, C., Wheeler, J. M., Choi, H.-J., & Cohen, A. S. (2021). Model selection for latent dirichlet allocation with small number of topics. In *Paper Presented at the Annual Meeting of the National Council on Measurement in Education (Virtual)*.
- Ponweiser, M. (2012). Latent dirichlet allocation in r.
- Schofield, A., Magnusson, M., & Mimno, D. (2017). Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (pp. 432–436).
- Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin, 24*(4), 35–43.
- Wheeler, J. M., Cohen, A. S., Xiong, J., Lee, J., & Choi, H.-J. (2021). Sample size for latent dirichlet allocation of constructed-response items. In M. Wiberg, D. Molenaar, J. González, U. Böckenholt, & J.-S. Kim (Eds.), *Quantitative Psychology: 85th Annual Meeting of the Psychometric Society* (pp. 263–273). Berlin: Springer.
- Wheeler, J. M., Engelhard, G., & Wang, J. (2022). Exploring Rater Accuracy Using Unfolding Models Combined with Topic Models Incorporating Supervised Latent Dirichlet Allocation. *Measurement: Interdisciplinary Research and Perspectives, 20*, 34–46.
- Xiong, J., Choi, H.-J., Kim, S., Kwak, M., & Cohen, A. S. (2019). Topic modeling of constructed-response answers on social study assessments. In M. Wiberg, D. Molenaar, J. González, U. Böckenholt, & J.-S. Kim (Eds.), *Quantitative Psychology: 84th The Annual Meeting of the Psychometric Society* (pp. 263–274).
- Xiong, J., Wheeler, J. M., Choi, H.-J., Lee, J., & Cohen, A. S. (2021). An empirical study of developing automated scoring engine using supervised latent dirichlet allocation. In M. Wiberg, D. Molenaar, J. González, U. Böckenholt, & J.-S. Kim (Eds.), *Quantitative Psychology: 85th Annual Meeting of the Psychometric Society* (pp. 429–438). Berlin: Springer.

Methods to Retrofit and Validate Q-Matrices for Cognitive Diagnostic Modeling



Charles Vincent Hunter , Hongli Li , and Ren Liu 

Abstract Cognitive diagnostic models (CDMs) are a family of constrained latent class models that estimate relationships between observed item responses and latent attributes (Rupp and Templin, *Educ Psychol Meas* 68:78–96, 2008). An important input needed in any CDM is the Q-matrix, an item-by-attribute table that represents a particular hypothesis about which attributes are required to answer each test item successfully. A large number of CDMs have been developed; however, many applications involve retrofitting a CDM to an existing non-diagnostic test. In this study, we conducted a systematic review to describe the current picture of retrofitting Q-matrices to non-diagnostic tests and consequently using the tests for diagnostic purposes.

Keywords CDM · Q-matrix · Retrofit

1 Introduction

Cognitive diagnostic models (CDMs) are a family of constrained latent class models that estimate relationships between observed item responses and latent traits (Rupp & Templin, 2008). These models assume that the items measure multiple latent traits and the latent traits are categorical (Liu & Shi, 2020). CDMs have been advocated as having the potential to provide rich diagnostic information from tests to aid

C. V. Hunter (✉)

Research, Evaluation, Assessment, and Accountability, Clayton County Public Schools,
Jonesboro, GA, USA

e-mail: charles.hunter@clayton.k12.ga.us

H. Li

Georgia State University, Atlanta, GA, USA

e-mail: hli24@gsu.edu

R. Liu

University of California Merced, Merced, CA, USA

e-mail: rliu45@ucmerced.edu

Table 1 Sample Q-matrix

	Attribute A	Attribute B	Attribute C
Item 1	1	0	0
Item 2	0	1	1
...

instruction and learning, because they provide a profile for each student in regard to whether or not the student has mastered the required skills (a.k.a. attributes) to provide correct responses to the test items. CDMs, therefore, are able to provide useful diagnostic feedback to teachers and students.

As summarized by DiBello et al. (2007), a systematic cognitive diagnostic assessment involves six steps: (i) describing assessment purpose; (ii) describing skill space; (iii) developing assessment tasks; (iv) specifying psychometric model; (v) performing model calibration and evaluation; and (vi) score reporting. However, very few large-scale tests are designed under a cognitive diagnostic modeling framework. Therefore, in most CDM applications, a non-diagnostic preexisting test is analyzed, which is referred to as retrofitting (Liu et al., 2017). A major challenge involved in retrofitting is that constructing the post-hoc Q-matrix is time consuming. In addition, calibrating an existing unidimensional test with a multidimensional CDM may not work or may be inefficient (Haberman & von Davier, 2007).

The Q-matrix that represents a particular hypothesis about which attributes are required to answer each test item successfully (Tatsuoka, 1983). As shown in Table 1, each row represents an item of the test and each column represents an attribute. A Q-matrix can have a simple structure—each item requires only one attribute—or a complex structure—at least one item requires more than one attribute (Rupp et al., 2010). A sound Q-matrix is critical for a successful CDM application (Gorin, 2009).

2 Method

Studies that met the following criteria were included in our review. First, the study had to apply CDM(s) to a real dataset. If a study adopted a Q-matrix developed and validated in a previous study, we would only keep the earlier study to avoid duplicates. For example, Jang et al. (2015) used the same Q-matrix that was developed and validated in Jang et al. (2013). We only included Jang et al. (2013). Second, only journal articles in English from 1983 to March 2021 were included.

To begin with, we performed a systematic search of ERIC and APA PsychInfo using keywords “cognitive diagnostic” or “diagnostic classification.” The authors read the full-text of each article to decide whether it was eligible. Then, we searched Google scholar using the same keywords “cognitive diagnostic” or “diagnostic classification.” The authors went through each entry to decide if any new study would be added to our previous findings. Finally, we consulted the studies included in Sessoms and Henson (2018), who reviewed the application of CDMs from 2008 to 2016.

After selecting the initial set of articles, the authors held several rounds of discussions among themselves to refine the selection criteria. During these discussions, we decided to restrict the study to journal articles published in English, in order to complete the study in a timely manner. We also realized that several articles used a Q-matrix developed in an earlier study, which would have created duplicates. Because of this, we decided to include only the earliest study. In the end, we found 80 articles that met our inclusion criteria. Seven of the articles reported two different CDM studies. For each of these articles, two studies were included for the final coding. Therefore, 87 studies from 80 published journal articles were included in this review.

We first drafted a coding sheet based on existing literature and prior CDM study review (e.g., Sessoms & Henson, 2018), with an emphasis on the Q-matrix development and validation. After several rounds of discussion and training of the coding procedure, one author coded all the studies. Then another author went through the coding of each article multiple times. All discrepancies and questions were resolved through discussion and negotiation until a consensus was reached.

3 Results

The earliest study included in our review was published in 1993 (Birenbaum & Tatsuoka, 1993). Fewer than 10 articles per year were published until 2019, when 11 were published. Seventeen were published in 2020. The articles were published in a variety of journals with journals focused on education and psychology being the most frequent categories, such as *Studies in Educational Evaluation* and *Frontiers in Psychology*. The most studies conducted in a single country were 34 in the US, followed by 11 in Iran. Thirteen studies used data collected from multiple countries, primarily because the tests used were the Programme for International Student Assessment (PISA) or the Trends in International Mathematics and Science Study (TIMSS). Many sample sizes were extremely large because of the multinational samples, and the Korean National Assessment of Educational Achievement (NAEA) exam, which had a sample of over 16 million examinees (Table 2). The number of items per assessment ranged from 4 to 216; and the number of attributes ranged from 2 to 27.

Table 2 Study descriptive statistics

	Sample size ^a	Sample size ^b	Number of items	Number of attributes
Max	16,928,895	120,767	216	27
Min	96	96	4	2
Mean	415,952.8	5449.0	39.3	8.3
Std dev	2,612,318.7	15,749.3	34.9	5.1
Median	1454	1252	33	7
Mode	10,000	10,000	20	5

^aWith Kim (2014)

^bWithout Kim (2014)

Table 3 Content area studied

Area studied	Frequency
Language skills ^a	30
Mathematics	30
Psychology (normal and pathological)	10
Listening	9
Science and medicine	2
Civic knowledge	1
Intrapreneurship	1
MOOC engagement	1
Professional competencies	1
Situational judgment	1
Social justice advocacy	1

^aReading, writing, grammar, foreign language skills, and foreign language arts

In terms of the construct being tested, language skills and mathematics, which appeared in 30 studies each, were the most frequent. Ten studies looked at different areas of psychology, such as personality, and pathological behavior (Table 3).

In all the studies, the attribute classification was dichotomous (i.e., master vs. non-master). Six of the studies reported correlations among the attributes. Although many of the studies discussed how to prepare score reports, none of the studies reported whether such diagnostic results were actually delivered to students or teachers.

The Q-matrices in 74 studies had a complex structure, while eight had a simple structure. In five of the studies, the relationships among the attributes were hierarchical. In 24 of the studies, the authors modified their initial Q-matrices, and in 52 of the studies, the authors provided the final Q-matrix. A variety of CDMs were used in the applications, and it was common for one study to apply multiple CDMs. The deterministic input, noisy “and” gate (DINA) and the generalized deterministic input, noisy “and” gate (GDINA) models were the most frequently used and appeared in 23 studies each, followed by the Reduced Reparameterized Unified Model (RRUM or FUSION) model (20 times; Table 4).

The Q-matrices were developed using combinations of different techniques (Table 5). The most common was using a literature review to determine the attributes (or skills) needed to respond to the items correctly. Review of the assessment items by content experts was the second most common method. Consulting the test specifications was used in eight studies, while asking examinees about how they answered questions was used in seven studies. Thirty-six studies did not report how the Q-matrix was developed.

Checking model fit indices was the most common means of validating the Q-matrix. Indices of both absolute fit (e.g., SRMSR, MAD) and relative fit (e.g., AIC, BIC) were used (Chen et al., 2013). Relative fit indices were used to compare different Q-matrices or different CDMs, and to compare CDM results to results from Classical Test Theory (CTT) or Item Response Theory (IRT) models. The Wald test

Table 4 CDM used

CDM	Number of articles
ACDM	6
Attribute hierarchy method (AHM)	3
DINA ^a	23
DINO ^b	13
GDINA ^c	23
GDM	4
Hierarchical diagnostic classification model (HDCM)	1
Log-linear cognitive diagnosis model (LCDM)	4
LLM	3
Mixture model	1
RRUM/FUSION ^d	20
Rule space method	10

^aIncludes four variant types of DINA

^bIncludes three Bayesian variant types of DINO

^cIncludes one variant type of GDINA

^dIncludes three variant types of RRUM

Table 5 Q-matrix development and validation

Development		Validation	
Task	Frequency	Task	Frequency
Literature review	39	Model fit indices	25
Not reported	36	Attribute mastery predictions	13
Content expert review	29	Compare with CTT/IRT	12
Author coding	15	Not reported	12
Test framework – specifications	8	Item parameters	11
Student reports	7	Reliability indices	10
		Empirical validation algorithms	8
		Compare different Q-matrices	6
		Cross validation with other criteria	5
		Student interviews	4
		Factor analysis	3
		Regression	3
		Discuss with experts	2
		Review of misfitting items	1

was also used to compare different models. Checking attribute mastery predictions (both for accuracy and for consistency) was the second most common validation technique. This was used in 13 studies. Other methods include evaluating item parameters and reliability indices, as well as using empirical validation algorithms (e.g., de la Torre and Chiu’s (2016) validation method, implemented in R package GDINA (Ma & de la Torre, 2020)). Twelve studies did not report their validation methods.

4 Discussion and Significance

The Q-matrix is of vital importance for the proper functioning of a CDM. If the Q-matrix is misspecified, the usefulness of the CDM is impaired (Gorin, 2009). It is, therefore, important to have a Q-matrix that is well-founded theoretically, as well as supported by empirical evidence (Rupp et al., 2010).

Developing a Q-matrix is an iterative process that involves theoretical guidance and content knowledge about the construct being tested. After an initial set of attributes has been developed, the Q-matrix needs to be refined to ensure that there are sufficient high-quality items for each attribute to produce stable results. Also, redundant or closely overlapped attributes need to be identified, combined or removed for parsimony. This is frequently done by removing an attribute that has high correlation with other attributes, or that has low correlation with item difficulty (Buck & Tatsuoka, 1998).

The most common methods of Q-matrix development are literature reviews and ratings by content experts. In addition, consulting test specifications and consulting with examinees via think-aloud or posttest interviews are good ways to understand cognitive processes when examinees respond to the test items. However, test specifications are often not available for a retrofitted test, and consulting examinees, which can provide valuable insights, sometimes may not be feasible given the development context. It is common to find a CDM study that uses multiple methods discussed above in their Q-matrix development stage (e.g., Li & Suen, 2013). Utilizing evidence from multiple sources greatly strengthens their Q-matrix.

Similarly, high quality CDM application studies tend to adopt multiple procedures to validate their Q-matrices from different perspectives. Our review shows that there are three main types of Q-matrix validation procedures: (a) comparing the CDM model results with results from CTT or IRT models; (b) examination of model fit indices; and (c) examination of attribute mastery predictions. First, it is not always valid to compare CDM results with results from IRT models. CDMs are multidimensional while IRT models are usually unidimensional. Even when multi-dimensional IRT is used, the assumptions are different as the latent variables in CDMs are categorical while the latent variables in IRT models are continuous. Therefore, such comparison does not always lead to meaningful results. Second, model fit indices have played an important role in Q-matrix validation. Both absolute and relative fit indices are utilized. With the lack of well-established criteria for the absolute fit indices for CDMs (Lei & Li, 2016), the relative fit indices (e.g., AIC, BIC) seem to be more useful when results from different models are compared. Some studies (e.g., Ravand et al., 2020) allowed the items to “choose” the best fitting model, but Hemati and Baghaei (2020) found that overall model fit for this procedure was not as good as using the GDINA model for all items.

Attribute mastery predictions usually consist of evaluating both classification accuracy and consistency of the whole latent class pattern for examinee responses. As Park et al. (2020) note, predicting the accuracy and consistency of examinee scores by a model is a measure of reliability. CDM studies in earlier years usually

did not report attribute reliability information, but more recent studies (Min & He, 2021; Wu et al., 2020, 2021) start to report such information.

In addition, in more recent years, a few studies (e.g., Effatpanah, 2019; Javidanmehr & Anani Sarab, 2019; Kilgus et al., 2020) used the Q-matrix empirical validation algorithm (de la Torre & Chiu, 2016) which was further available in the GDINA R package (Ma & de la Torre, 2020). This offers the possibility of a convenient way to validate the Q-matrix empirically. However, this empirical algorithm can only serve as a supplementary information for Q-matrix validation. As recommended by de la Torre (2008), it is always important to combine the Q-matrix empirical validation results with content knowledge.

Suggestions. Our findings suggest several procedures that should be followed when developing a retrofitted Q-matrix, as well as some procedures that should be used only with caution. Our primary recommendation is that researchers and practitioners should consider perspectives from both construct theory and statistical analysis. The process of developing a retrofitted Q-matrix should always include subject matter experts who know both theory and content of the test construct (Rupp et al., 2010). Second, the set of attributes developed should be based on the principle of parsimony where highly correlated attributes may be combined (Buck & Tatsuoka, 1998). Finally, more than one method needs to be used to develop the Q-matrix so that evidence from multiple sources can be combined to strengthen the validity of the Q-matrix (Li & Suen, 2013).

Once the Q-matrix has been developed statistical testing needs to be done to verify the appropriateness of the matrix. This can be done by testing for model fit and reliability using actual data, using both absolute and relative fit indices to compare different models (Lei & Li, 2016) to select the best fitting one. Also, researchers should test for reliability by evaluating both classification accuracy and classification consistency of the whole latent class pattern for examinee responses (Park et al., 2020). An empirical validation algorithm (e.g., de la Torre and Chiu's (2016) empirical validation model for DINA) could also be used.

We recommend against comparing results from CDM models with IRT or CTT models, because they are based on different theory and are not strictly comparable. Results from such comparisons may not be meaningful. For optimal model fit, given sufficient sample size, we recommend starting the analysis with a saturated CDM and examining the significance of the main effects and interaction effects (if any), before considering specific smaller CDMs with particular assumptions on the relationship between items and attributes (Hemati & Baghaei, 2020).

Limitations. A major limitation of this review is that we only included journal articles published in English. Adding dissertations, conference presentations, and articles published in other languages has the potential of opening up more methods of Q-matrix development and validation, as well as insights into the CDM applications. These are areas for continued work. Furthermore, some CDM studies did not provide details about their Q-matrix development and validation procedures so that we were not able to code such information for every study included in the review. We, therefore, call for a detailed report of Q-matrix development and validation procedures in future CDM application studies.

Notwithstanding these limitations, this review contributes to the research into Q-matrix and CDM applications by highlighting the present state of Q-matrix development and validation, some of the possible tools for the process, and the need to use multiple methods in developing and validating Q-matrices.

References

- Birenbaum, M., & Tatsuoka, K. K. (1993). Applying an IRT-based cognitive diagnostic model to diagnose students' knowledge states in multiplication and division with exponents. *Applied Measurement in Education*, 6, 255–268. https://doi.org/10.1207/s15324818ame0604_1
- Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, 15, 119–157. <https://doi.org/10.1191/026553298667688289>
- 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. <https://doi.org/10.1111/j.1745-3984.2012.00185.x>
- de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, 45, 343–362. <https://doi.org/10.1111/j.1745-3984.2008.00069.x>
- de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, 81, 253–273. <https://doi.org/10.1007/s11336-015-9467-8>
- DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 979–1030). Elsevier. [https://doi.org/10.1016/S0169-7161\(06\)26031-0](https://doi.org/10.1016/S0169-7161(06)26031-0)
- Effatpanah, F. (2019). Application of cognitive diagnostic models to the listening section of the International English Language Testing System (IELTS). *International Journal of Language Testing*, 9, 1–28. https://www.ijlt.ir/?action=article&au=801055&_au=Effatpanah,%20Farshad
- Gorin, J. S. (2009). Diagnostic classification models: Are they necessary? Commentary on Rupp and Templin (2008). *Measurement: Interdisciplinary Research and Perspectives*, 7, 30–33. <https://doi.org/10.1080/15366360802715387>
- Haberman, S. J., & von Davier, M. (2007). Some notes on models for cognitively based skills diagnosis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 1031–1038). Elsevier. [https://doi.org/10.1016/S0169-7161\(06\)26040-1](https://doi.org/10.1016/S0169-7161(06)26040-1)
- Hemati, S. J., & Baghaei, P. (2020). A cognitive diagnostic modeling analysis of the English Reading Comprehension section of the Iranian National University Entrance Examination. *International Journal of Language Testing*, 10, 11–32. <https://eric.ed.gov/?id=EJ1291043>
- Jang, E. E., Dunlop, M., Wagner, M., Kim, Y. H., & Gu, Z. (2013). Elementary school ELLs' reading skill profiles using cognitive diagnosis modeling: Roles of length of residence and home language environment. *Language Learning*, 63, 400–436. <https://doi.org/10.1111/lang.12016>
- Jang, E. E., Dunlop, M., Park, G., & van der Boom, E. H. (2015). How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback? *Language Testing*, 32, 359–383. <https://doi.org/10.1177/0265532215570924>
- Javidanmehr, Z., & Anani Sarab, M. R. (2019). Retrofitting non-diagnostic reading comprehension assessment: Application of the G-DINA model to a high stakes reading comprehension test. *Language Assessment Quarterly*, 16, 294–311. <https://doi.org/10.1080/15434303.2019.1654479>

- Kilgus, S. P., Bonifay, W. E., Eklund, K., von der Embse, N. P., Peet, C., Izumi, J., Shim, H., & Meyer, L. N. (2020). Development and validation of the Intervention Skills Profile–Skills: A brief measure of student social-emotional and academic enabling skills. *Journal of School Psychology, 83*, 66–88. <https://doi.org/10.1016/j.jsp.2020.10.001>
- Kim, H. (2014). Application of cognitive diagnostic model for achievement profile analysis. *KAERA Research Forum, 1*(1), 15–25.
- Lei, P.-W., & Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and Q-matrices. *Applied Psychological Measurement, 40*, 405–417. <https://doi.org/10.1177/0146621616647954>
- Li, H., & Suen, H. K. (2013). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment, 18*(1), 1–25. <https://doi.org/10.1080/10627197.2013.761522>
- Liu, R., & Shi, D. (2020). Using diagnostic classification models in psychological rating scales. *The Quantitative Methods for Psychology, 16*, 442–456. <https://doi.org/10.20982/tqmp.16.5.p442>
- Liu, R., Huggins-Manley, A. C., & Bradshaw, L. (2017). The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educational and Psychological Measurement, 77*, 220–240. <https://doi.org/10.1177/0013164416645636>
- Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software, 93*(14), 1–26. <https://doi.org/10.18637/jss.v093.i14>
- Min, S., & He, L. (2021). Developing individualized feedback for listening assessment: Combining standard setting and cognitive diagnostic assessment approaches. *Language Testing, 39*(1), 90–116. <https://doi.org/10.18637/jss.v093.i14>
- Park, Y. S., Morales, A., Ross, L., & Paniagua, M. (2020). Reporting subscore profiles using diagnostic classification models in health professions education. *Evaluation & the Health Professions, 43*, 149–158. <https://doi.org/10.1177/0163278719871090>
- Ravand, H., Baghaei, P., & Doebler, P. (2020). Examining parameter invariance in a general diagnostic classification model. *Frontiers in Psychology, 10*, 2930. <https://doi.org/10.3389/fpsyg.2019.02930>
- Rupp, A. A., & Templin, J. (2008). The effects of q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*, 78–96. <https://doi.org/10.1177/0013164407301545>
- Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford.
- Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives, 16*, 1–17. <https://doi.org/10.1080/15366367.2018.1435104>
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354. <https://doi.org/10.1111/j.1745-3984.1983.tb00212.x>
- Wu, X., Wu, R., Chang, H.-H., Kong, Q., & Zhang, Y. (2020). International comparative study on PISA Mathematics Achievement Test based on cognitive diagnostic models. *Frontiers in Psychology, 11*, 1–13. <https://doi.org/10.3389/fpsyg.2020.02230>
- Wu, X., Zhang, Y., Wu, R., & Chang, H. H. (2021). A comparative study on cognitive diagnostic assessment of mathematical key competencies and learning trajectories. *Current Psychology, 1*–13. <https://doi.org/10.1007/s12144-020-01230-0>

The Sum Scores and Discretization of Variables Under the Linear Normal One-Factor Model



Rudy Ligтвоet

Abstract The sum score is often used in practical test applications and the joining of outcomes is common practice, when preparing response data for analysis. Yet, many models for response data are not designed for this kind of handling of data. Research on the use of the sum score for stochastic inferences and the discretization of response variables is extended to the linear normal one-factor. It is shown that the model implies a stochastic ordering on the latent factor by the sum of the observed variables, but that this property no longer needs to hold when variables are discretized prior to taking the sum score. The implications of this result are discussed.

Keywords Discretization · Linear normal one-factor model · Pólya frequency functions · Sum scores · Totally positive densities

1 Introduction

For test and questionnaire data, respondents are often assigned a latent value for the attribute that the test is aimed to measure, based on a model for the dependencies that exist between the test items. Because the latent variable is unobserved, it is convenient to consider the observable sum scores across the test items as a proxy for the latent values instead. For the use of the sum score, a desirable feature of a measurement model would then be that a higher sum score also corresponds to a higher expected latent value, so that the ordering of respondents by their sum scores *stochastically* agrees with the ordering by their latent values. The use of the sum score for making such ordinal inferences has been studied for various item response theory models (Hemker et al., 1996, 1997; Ligтвоet, 2012, 2015), based on a *monotone likelihood ratio* (MLR) ordering of the latent variable by the sum scores.

R. Ligтвоet (✉)
University of Cologne, Cologne, Germany
e-mail: rlichtvoe@uni-koeln.de

For these models, Hemker (2001) also looked at the effect of joining response categories. The results show that the MLR property is not implied by all models, and that for some models the model may be invalidated when (adjacent) outcomes are joined (cf. Andrich, 1995a, 1995b; Roskam, 1995). With the omnipresent use of the sum (or average) score across many test applications and the common practice of collapsing outcomes (e.g., median split), there thus seems to be a mismatch between the models proposed for response data and the way these data are handled in practice. The present chapter looks at the use of the sum score and the discretization of response data under the *linear normal one-factor* (LNF) model.

1.1 The Linear Normal One-Factor

Factor analysis provides a framework to account for dependencies that exist between the multiple item response variables of a test, whereby item response variables serve as indicators of the common factors that the test aims to measure. The LNF model proposes a single latent variable or factor Z to account for the covariances between the item variables X_1, \dots, X_n by the linear relationships

$$X_i = a_i Z + U_i, \text{ with } a_i > 0,$$

where a_i denotes the i th factor loading and U_i is the i th residual or unique factor. It is assumed that U_1, \dots, U_n, Z are independent and normally distributed, with zero means (centered), and (non-negative) variances

$$\text{Var}(U_i) = \sigma_i^2 \text{ and } \text{Var}(Z) = \sigma^2.$$

Further, assume that $\text{Cov}(U_i, U_j) = 0$ (for $i \neq j$) and $\text{Cov}(U_i, Z) = 0$ (Jöreskog, 1971; Lord & Novick, 1968). Hence, under the LNF model, the variables X_1, \dots, X_n are *conditionally independent* (CI), given $Z = z$.

1.2 Monotone Transformations

In this chapter, two monotone (non-decreasing) transformations are considered that are often used on X_1, \dots, X_n in practice. Let $\mathbf{X} = (X_1, \dots, X_n)$ denote the random vector containing the variables X_i , with realizations $\mathbf{x} \in \mathbb{R}^n$. Then, a function $\phi(\mathbf{x})$ is said to be monotone, whenever $\mathbf{x} < \mathbf{y}$ (element-wise) implies that $\phi(\mathbf{x}) \leq \phi(\mathbf{y})$.

The Sum Score The first transformation that is considered is the sum score $S = X_1 + \dots + X_n$ often used in practice as a proxy for Z (McNeish & Wolf, 2020), with $\phi(\mathbf{x})$ representing a mapping of many-to-one or aggregation; i.e., $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$. In practice, a LNF model is often fitted to the data in order to assess the validity

(factorial structure) of the test, whereas in subsequent analysis the sum or average score is used to ascertain the validity indices (e.g., predictive validity) of classical test theory. A minimal requirement of the LNF model for such practice is that the model implies a stochastic ordering of the factor scores by the sum scores, such that higher factor scores are expected for higher sum scores. Let $f(s, z)$ denote the joint density of S and Z , then the stochastic ordering by the sum scores is satisfied, whenever the density $f(s, z)$ is *totally positive of order 2*; that is

$$f(s_1, z_1)f(s_2, z_2) \geq f(s_1, z_2)f(s_2, z_1), \quad (\text{TP}_2)$$

for all $s_1 < s_2$ and $z_1 < z_2$ (Karlin, 1968, cf. the MLR property).

Discretization As a second transformation, consider the discretization of the variables X_1, \dots, X_n . The LNF model is often used for the analysis of discrete ordinal response data (Jöreskog and Moustaki, 2001), where X_1, \dots, X_n are taken as the ghosts underlying the observable discrete variables V_1, \dots, V_n . Here, $\phi(\mathbf{x})$ takes on the form $(\phi_1(x_1), \dots, \phi_n(x_n))$, with

$$V_i = \phi_i(X_i; b_1, \dots, b_{m_i}) \text{ and } b_1 < \dots < b_{m_i},$$

where

$$V_i = v_i, \text{ whenever } b_v \leq x_i < b_{v+1}$$

($b_{m_i+1} = -b_0 = \infty$ by definition). In words, each discretization ϕ_i proposes m_i ordered thresholds, where v_i denote the largest threshold passed by the outcome of X_i . With $v_i \in \{0, 1, \dots, m_i\}$, the outcomes of V_i are said to be *equidistant* (Andrich, 1995a). Sijtsma and Van der Ark (2017) discuss the use of an equidistant scoring rule in relationship with the use of the sum score in the context of Mokken's *monotone homogeneity* (MH) model (Mokken, 1971; Molenaar, 1997). In addition to CI and the *unidimensionality* assumption, the MH model assumes that the tail distributions $1 - F(x_i|z)$ are non-decreasing in z (Holland & Rosenbaum, 1986). The LNF model satisfies the MH model assumptions, also after applying a discretization to X_1, \dots, X_n . In case $m_i = m = 1$, the transformation $\phi(\mathbf{x})$ corresponds to a dichotomization of the response variables.

For later reference, the concept of *Pólya frequency functions of order 2* (PF₂) is introduced (e.g., Efron, 1965; Schoenberg, 1951).

Definition 1 The density $f(x)$ is said to be PF₂, if for all $x_1 < x_2$ and $y_1 < y_2$

$$f(x_1 - y_1)f(x_2 - y_2) \geq f(x_1 - y_2)f(x_2 - y_1). \quad (1)$$

Ellis (2015) showed that the (monotone higher-order) one-factor model, with residuals having PF₂ densities (e.g., normally distributed), implies that $f(\mathbf{v})$ is

multivariate TP_2 (Karlin & Rinott, 1980). This in turn implies that each $f(v_i, v_j)$ is TP_2 .

Strictly speaking, the normality requirements of the LNF model does not hold for the discrete variables V_1, \dots, V_n . However, the LNF may still provide an adequate approximation of discrete response data (Rhemtulla et al., 2012).

Chapter Overview The purpose of this chapter is to investigate the effect the discretization of the indicators X_1, \dots, X_n has on the use of the sum score for the stochastic ordering on Z . In the next section, it is shown that the LNF model implies a stochastic ordering on Z by the sum score S . However, it is also shown that the stochastic ordering property does not generally hold when the sum score is used after discretizing the indicators X_1, \dots, X_n , except in the special case of a dichotomization. These results and their implications are further discussed in Sect. 3.

2 The LNF Model and the Sum Score

In this section, it is shown that the LNF model implies a stochastic ordering on Z by the sum $S = X_1 + \dots + X_n$. However, after discretization of the variables X_1, \dots, X_n obtained under the LNF model, the sum $R = V_1 + \dots + V_n$ of the newly obtained variables V_1, \dots, V_n no longer needs to provide a stochastic ordering on the factor Z . That is, $f(s, z)$ is TP_2 does not imply that $f(r, z)$ is also TP_2 .

2.1 Preliminaries

In order to show that the LNF model implies a stochastic ordering of Z by $S = X_1 + \dots + X_n$, it is convenient to express the model in terms of the more general properties TP_2 and PF_2 . To this end, consider the joint and conditional densities $f(x_i, z)$ and $f(x_i|z)$, respectively. First, assuming CI. Then, with $E(X_i) = E(Z) = 0$, the covariance between X_i and Z equals

$$E(X_i Z) = a_i E(Z^2) + E(U_i Z) = a_i \sigma^2 > 0.$$

This implies that $f(x_i, z)$ is TP_2 (Karlin & Rinott, 1983). Second, because U_i is normally distributed, so is the conditional density $f(x_i|z)$. Consequently, $f(x_i|z)$ has a PF_2 density (Efron, 1965). Note that, for strictly positive densities, if $f(x_i, z)$ is TP_2 , then $f(x_i|z)$ is TP_2 as a function of (x_i, z) (Holland & Rosenbaum, 1986). Because the TP_2 property implies that the tail distribution $1 - F(x_i|z)$ is non-decreasing in z , it thus follows that the LNF model satisfies the assumptions (i.e., is a special case) of the MH model (Holland & Rosenbaum, 1986).

The following observation is useful for the proof of Theorem 1 below. Assume that $f(x, y|z) > 0$ (as implied by the LNF model). Then, for $z_1 < z_2$, the inequality

$$f(x_2, y_2|z_2)f(x_1, y_1|z_1) \geq f(x_2, y_2|z_1)f(x_1, y_1|z_2) \tag{2}$$

holds, whenever

$$\frac{f(x_2, y_2|z_2)}{f(x_2, y_2|z_1)} \geq \frac{f(x_2, y_1|z_2)}{f(x_2, y_1|z_1)} \geq \frac{f(x_1, y_1|z_2)}{f(x_1, y_1|z_1)}.$$

Hence, (2) holds, if both (a) $f(x, y|z)$ is TP₂ as a function of (y, z) , with $y_1 < y_2$ and $X = x_2$ (fixed), and (b) $f(x, y|z)$ is TP₂ in (x, y) , with $x_1 < x_2$ and $Y = y_1$.

The next result is proven in Ligotvoet (2021), but here adapted to the LNF model.

Theorem 1 *The LNF model implies that $f(s, z)$ is TP₂.*

Proof Suppose that Theorem 1 holds for $n = 2$, with $X_1 = X$ and $X_2 = Y$. The proof of Theorem 1 then follows, by sequentially taking $X = X_1 + \dots + X_{i-1}$ and $Y = X_i$, for $i = 2, \dots, n$. Hence, it is sufficient to show that $f(s, z)$ is TP₂, for $S = X + Y$.

Due to CI, the conditional density of S is given by the convolution $f(s|z) = \int g(x|z)h(s - x|z)dx$. Then, $f(s, z)$ is TP₂, if for any $s_1 < s_2$ and $z_1 < z_2$ it holds that $f(s_2|z_2)f(s_1|z_1) \geq f(s_2|z_1)f(s_1|z_2)$, which yields

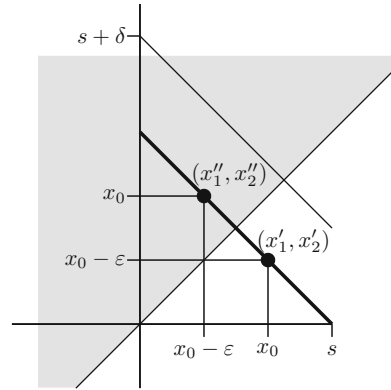
$$\int_{-\infty}^{s_1} \int_{-\infty}^{s_2} g(x_2|z_2)h(s_2 - x_2|z_2)g(x_1|z_1)h(s_1 - x_1|z_1) - g(x_2|z_1)h(s_2 - x_2|z_1)g(x_1|z_2)h(s_1 - x_1|z_2) dx_2dx_1 \geq 0. \tag{3}$$

Note that (3) has the form of (2), with $f(x, y|z) = g(x|z)h(s - x|z)$. So, it is sufficient to show that (3) holds for the case that X is constant between z_1 and z_2 , and the case that Y is constant between z_1 and z_2 . Because both cases are symmetric in their arguments, we'll only consider taking Y to be constant at $Z = z$ (for $z_1 \leq z \leq z_2$). For (3), this yields

$$\int_{-\infty}^{s_1} \int_{-\infty}^{s_2} h(s_2 - x_2|z)h(s_1 - x_1|z) \cdot (g(x_2|z_2)g(x_1|z_1) - g(x_2|z_1)g(x_1|z_2)) dx_2dx_1 \geq 0. \tag{4}$$

Because $g(x, z)$ is TP₂, the function within the integral of (4) has a positive outcome for $x_1 < x_2$ and negative values for $x_1 > x_2$. What remains to be shown is that the density of the area of the integral spanning all $x'_1 > x'_2$ is smaller than the area spanning all $x''_1 < x''_2$ (see Fig. 1 for illustration). Let $x'_1 = x_0$ and $x'_2 = x_0 - \varepsilon$, with $\varepsilon > 0$, and accordingly $x''_1 = x_0 - \varepsilon$ and $x''_2 = x_0$. This yields a one-to-one (injective) mapping of each pair (x'_1, x'_2) that yields negative values in (4) to (x''_1, x''_2) , as shown in Fig. 1. Also, let $s_1 = s$ and $s_2 = s + \delta$, with $\delta > 0$. Then, it is

Fig. 1 Plot of x_1 and x_2 , with the point (x'_1, x'_2) in the gray areas above the identity yielding positive values of (4)



sufficient to show that for all values $x_0, \varepsilon, s, \delta$, and $z_1 < z_2$,

$$(g(x_0|z_2)g(x_0 - \varepsilon|z_1) - g(x_0|z_1)g(x_0 - \varepsilon|z_2)) \cdot$$

$$(h(s - x_0 + \varepsilon|z)h(s - x_0 + \delta|z) - h(s - x_0 + \varepsilon + \delta|z)h(s - x_0|z)) \geq 0.$$

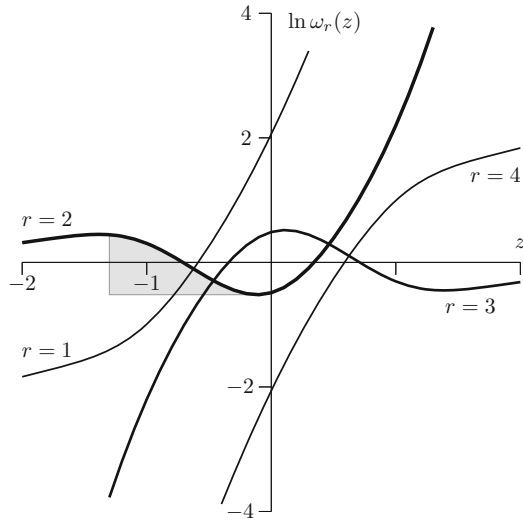
The first part of (5) in parentheses is non-negative, because $g(x, z)$ is TP₂. The second part in parentheses is also non-negative, because $h(y|z)$ is PF₂. This can be seen by taking $x_1 = s - x_0, x_2 = s - x_0 + \varepsilon, y_1 = -\delta$, and $y_2 = 0$ in (1). □

2.2 The Sum Score After Discretization of Variables

Theorem 1 shows that under the LNF model $f(s, z)$ is TP₂ for the sum score $S = X_1 + \dots + X_n$. Next, a discretization is considered, whereby we denote the sum of the discretized variables as $R = V_1 + \dots + V_n$. As mentioned earlier, the LNF model satisfies the MH model assumptions. For the special case when a dichotomization is applied to all variables (i.e., $m_i = m = 1$), the LNF reduces to Mokken’s MH model for binary variables, which has been shown to imply a stochastic ordering of the latent variable by the sum score (Ghurye & Wallace, 1959; Grayson, 1988; Huynh, 1994; Ünlü, 2008). For $m_i > 2$, however, the MH model does not imply a stochastic ordering of the latent variable by the sum score (Hemker et al., 1996, 1997). The next example also shows that $f(r, z)$ need not be TP₂, after discretizing the variables X_1, \dots, X_n obtained under the LNF model.

Example 1 Consider the LNF for $n = 2$ variables, with $a_1 = a_2 = 1, \sigma^2 = 1$, and $\sigma_1^2 = 2/5$ and $\sigma_2^2 = 5/2$. Also, let $V_i = \phi_i(X_i; -1/2, 1/2)$, for $i = 1, 2$ (i.e., $m_i = m = 2$), and $R = V_1 + V_2$. Further, define the log-odds $\ln \omega_r(z) = \ln f(r|z) - \ln f(r - 1|z)$, for $r = 1, \dots, 4$, which are non-decreasing in z , whenever $f(r, z)$ is TP₂. Figure 2 shows that for $r = 2, 3$, the log-odds are decreasing (in violation of the stochastic property). For $r = 2$, the gray area in Fig. 2 shows the decrease in

Fig. 2 Log-odds for the sum scores as a function of z , showing a violation (decrease; e.g., gray area) of the stochastic ordering of Z by $R = V_1 + V_2$



log-odds between $z = -1.37$ and $z = -0.10$, indicating that for two subjects with these factor scores, the subject with the (higher) factor score $z = -0.10$ is about 2.5 times more likely to obtain the lower sum score than the subject that has a factor score that is more than one standard deviation lower (i.e., a substantial violation).

3 Discussion

The property $f(s, z)$ is TP₂ was proposed as a minimal requirement for the use of the sum $S = X_1 + \dots + X_n$. This property is less restrictive than the tau-equivalence requirement from classical test theory (Lord & Novick, 1968; McNeish & Wolf, 2020), but is also limited to ordinal inferences. Theorem 1 implies that the confirmation of the LNF model justifies ordinal inferences about the latent factor based on the sum score. For applications that require more than mere ordinal inferences, the use of the estimated factor score may be more advantageous.

The discretization of variables obtained under the LNF model, not only jeopardizes the normality assumption, but also has implications for the practical use of the sum score. The extent to which the stochastic ordering property by the sum score is violated, in a practical sense, will depend on the number of items variables of the test, as well as the number of categories resulting from the discretization. Simulation studies may further address this issue.

Instead of assuming a normal distribution to underlie the observed discrete ordinal response data, alternative approaches for analyzing these data impose restrictions on the cumulative distributions similar to Samejima’s (1969) *graded*

response model. Jöreskog and Moustaki (2001) and Takane and De Leeuw (1987) showed that the normal ogive model for graded responses is formally equivalent to the LNF model that assumes a normal distribution to underlie the ordinal responses. The difference between these models is that, for the graded response model, the conditional response distributions is discretized prior to taking the marginal across the latent factor, whereas for the LNF model the discretization is applied (afterwards) to the marginal distribution (Takane & De Leeuw, 1987, p. 397). In the latter case, the discretization may invalidate the property $f(s, z)$ is TP₂, when applying the LNF to discrete ordinal response data. That the graded response model does not imply this property was already shown by Hemker et al. (1996, 1997).

To conclude, the use of the sum score, albeit practical, is not what most models are designed for. The applied researcher should realize that an ordering on a latent variable by the sum score is not something that can be simply assumed to hold. If the applied researcher has a model that accurately describes the response data, it might generally be best to rely on the model estimates, rather than using the sum scores. And if a transformation of the data is deemed necessary, the validity of the model will need to be reassessed.

References

- Andrich, D. (1995a). Models for measurement, precision, and the nondichotomization of graded responses. *Psychometrika*, *60*(1), 7–26.
- Andrich, D. (1995b). Further remarks on nondichotomization of graded responses. *Psychometrika*, *60*(1), 37–46.
- Efron, B. (1965). Increasing properties of Pólya frequency functions. *The Annals of Mathematical Statistics*, *36*(1), 272–279.
- Ellis, J. L. (2015). MTP2 and partial correlations in monotone higher-order factor models. In R. E. Millsap, D. M. Bolt, L. A. van der Ark, W. C. Wang (Eds.), *Quantitative psychology research* (pp. 261–272). Berlin: Springer.
- Ghurye, S. G., & Wallace, D. L. (1959). A convolutive class of monotone likelihood ratio families. *The Annals of Mathematical Statistics*, *30*(4), 1158–1164.
- Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika*, *53*(3), 383–392.
- Hemker, B. T. (2001). Reversibility revisited and other comparisons of three types of polytomous IRT models. In A. Boomsma, M. A. J. Van Duijn, T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 277–296). Berlin: Springer.
- Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika*, *61*(4), 679–693.
- Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, *62*(3), 331–347.
- Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, *14*(4), 1523–1543.
- Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. *Psychometrika*, *59*(1), 77–79.
- Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*(2), 109–133.

- Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, *36*(3), 347–387.
- Karlin, S. (1968). *Total positivity*. Redwood City: Stanford University Press.
- Karlin, S., & Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities, I. Multivariate totally positive distributions. *Journal of Multivariate Analysis*, *10*(4), 467–498.
- Karlin, S., & Rinott, Y. (1983). M-matrices as covariance matrices of multinormal distributions. *Linear Algebra and its Applications*, *52–53*, 419–438.
- Ligtvoet, R. (2012). An isotonic partial credit model for ordering subjects on the basis of their sum scores. *Psychometrika*, *77*(3), 479–494.
- Ligtvoet, R. (2015). A test for using the sum score to obtain a stochastic ordering of subjects. *Journal of Multivariate Analysis*, *133*, 136–139.
- Ligtvoet, R. (2021). Conditional TP₂ distributions of sums for latent ordinal inferences. Manuscript submitted for publication.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Boston: Addison-Wesley.
- McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*(6), 2287–2305.
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Berlin: De Gruyter.
- Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In *Handbook of modern item response theory* (pp. 369–380). Berlin: Springer.
- 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.
- Roskam, E. E. (1995). Graded responses and joining categories: A rejoinder to Andrich's "Models for measurement, precision and non-dichotomization of graded responses". *Psychometrika*, *60*(1), 27–35.
- Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika*, *34*, 1–97
- Schoenberg, I. J. (1951). On pólya frequency functions. I. The totally positive functions and their Laplace transforms. *Journal d'Analyse Mathématique*, *1*(1), 331–374.
- Sijtsma, K., & Van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 137–158.
- Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408.
- Ünlü, A. (2008). A note on monotone likelihood ratio of the total score variable in unidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 179–187.

Identifying Zones of Targeted Feedback with a Hyperbolic Cosine Model



Ye Yuan and George Engelhard

Abstract Formative assessments are used to identify student strengths and weaknesses, but they frequently do not identify targeted feedback strategies for improving student achievement. The study introduces the concept of Zones of Targeted Feedback for identifying the sets of optimal feedback strategies that can be used for improving student writing. The study suggests a non-cumulative responses process for linking feedback strategies and achievement levels. We offer an unfolding model as an alternative measurement paradigm to identify and improve the effectiveness of teacher feedback strategies linked to levels of student achievement. The study presents two examples (an illustrative one and an empirical one) of using an unfolding model called the Hyperbolic Cosine model to illustrate our conceptual framework.

Keywords Writing · Feedback · Unfolding model · Formative assessment

1 Introduction

Feedback is one of the most powerful influences on student achievement, and feedback can be viewed as a “consequence of performance” (Hattie & Timperley, 2007, p. 81). A significant body of research stresses the importance of feedback effectiveness in assessment and instruction (Bangert-Drowns et al., 1991; Kluger & DeNisi, 1996; Kulhavy, 1977; Kulhavy & Stock, 1989; Narciss & Huth, 2004; Shute, 2008; Willingham, 1990). Important tasks for teachers include the identification of a student’s level of achievement, and then selection of feedback strategies that can move the student forward in their learning. Feedback strategies on different areas may not be appropriate for a given student based on their current level of achievement. Therefore, educators must target their feedback for each student.

Y. Yuan (✉) · G. Engelhard
Department of Educational Psychology, University of Georgia, Athens, GA, USA
e-mail: ye.yuan@uga.edu; gengelh@uga.edu

The purposes of the study are to explore the use of an unfolding model to identify the Zones of Targeted Feedback, and to discuss how teachers might use Zones of Targeted Feedback to provide scaffolding to assist in the development of appropriate feedback strategies. The study first introduces the concept of Zones of Targeted Feedback, and then describes an unfolding model that can be used to link levels of student achievement with recommended feedback strategies. Next, the approach is illustrated with two examples. The first example is an illustration of the idea, and the second example is an application of our idea with an empirical study in the context of writing assessment. The results of this empirical study are briefly described.

2 Zones of Targeted Feedback

Feedback holds promise for improving student achievement, but the promise is not guaranteed and depends in no small part on the care taken in choosing an appropriate mode of feedback (Hattie & Timperley, 2007). Appropriate feedback can be related to the concept of Zone of Proximal Development (ZPD; Vygotsky, 1978). Vygotsky defined ZPD as “the distance between the actual developmental level as determined by independent problem-solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers” (1978, p. 86). In this study, we propose adapting this idea to identify zones of targeted feedback (ZTF) that can assist teachers in identifying effective feedback strategies based on student levels of achievement. Formative assessments provide the identification of achievement levels, while the method proposed in this study can be used to identify the targeted feedback strategies to match an individual student’s ZPD with the goal of improving each student’s level of achievement.

Figure 1 shows the conceptual map of the study. The achievement levels of three students (Students A, B, and C) are shown on the first line from low to high achievement. The second line defines the locations of feedback strategies (S1, S2, and S3). The ZTF can be defined based on the recommended feedback strategies identified by teachers for students located at different levels on the achievement continuum. Take student B as an example, the bell curve in the middle can be recognized as the ZTF for student B. Comparing other feedback on the second scale, the feedback strategies in this range are judged to be relatively effective and appropriate for student B. The optimal feedback for student B is the “peak” of the ZTF, which is feedback strategy S2. Teachers are less likely to provide the other two feedback strategies (S1, S3) as they are not in student B’s ZTF. In the context of writing, a feedback strategy *below* a student’s level might be related to feedback on the neatness of their handwriting, while a feedback strategy *above* a student’s proficiency level might focus on more complex matters of organization, such as transitions between sentences. The key idea is that feedback strategies below or above a student’s achievement level may not provide optimal feedback for improving student writing, so it is important to connect student achievement levels to a set of feedback strategies that define the ZTF for each student.

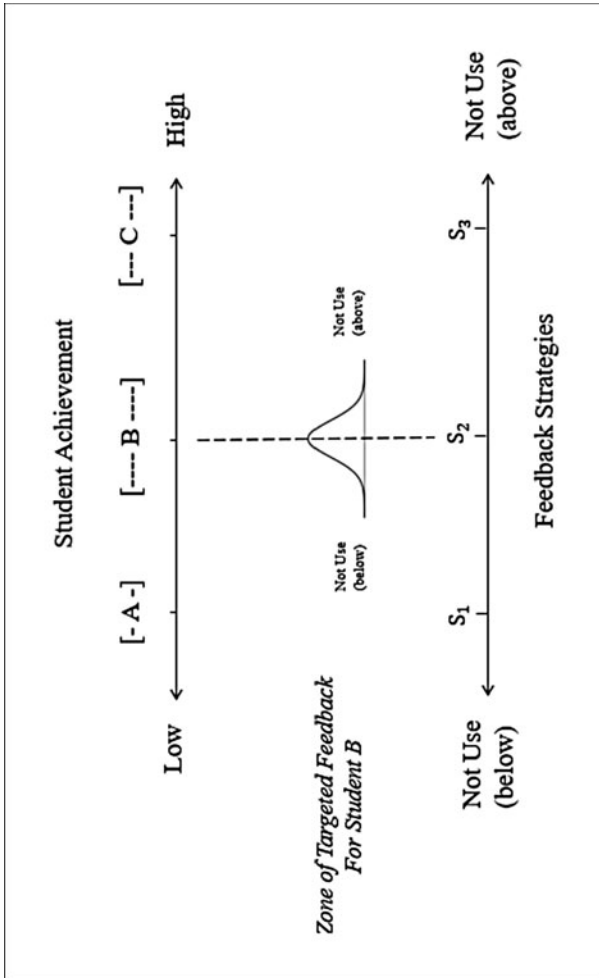


Fig. 1 Conceptual map of connection between levels of student achievement and zones of targeted feedback

Table 1 Ideal response pattern for cumulative and unfolding scales

Panel A Ideal Response Patterns														
Person	Cumulative Scale: Scalogram (Guttman, 1950)							Unfolding Scale: Parallelogram (Coombs, 1964)						
	A	B	C	D	E	F	G	A	B	C	D	E	F	G
1	1	0	0	0	0	0	0	1	1	0	0	0	0	0
2	1	1	0	0	0	0	0	1	1	1	0	0	0	0
3	1	1	1	0	0	0	0	0	1	1	1	0	0	0
4	1	1	1	1	0	0	0	0	0	1	1	1	0	0
5	1	1	1	1	1	0	0	0	0	0	1	1	1	0
6	1	1	1	1	1	1	0	0	0	0	0	1	1	1
7	1	1	1	1	1	1	1	0	0	0	0	0	1	1

Panel B Operating Characteristic Functions	
Deterministic	Deterministic

The study suggests viewing this bell curve of ZTF as an unfolding response process because the probability of a positive response is a single-peaked function (Andrich, 1997), which is different from cumulative response processes. Most measurement models are based on cumulative response processes. In the cumulative response processes, the probability of a positive response is a monotonic function of the relevant parameters. A comparison of a cumulative scale and an unfolding scale is shown in Table 1 (Andrich, 1988). Panel A in Table 1 illustrates a cumulative scale for seven persons and seven items (A–G). The persons and items are ordered by row and column scores. This ordering yields the distinctive triangular pattern that defines a Guttman scale (positive responses highlighted). The unfolding scale is also ordered, and the responses exhibit a parallelogram structure that is iconic for an unfolding scale (positive responses highlighted). Panel B in Table 1 shows the underlying operating functions for the two scales. A Guttman scale has a distinctive staircase pattern, while the unfolding scale can be represented by distinctive top-hat pattern. The comparison helps to illustrate the distinction between the cumulative and unfolding principles.

3 Hyperbolic Cosine Model

This study proposes to use a Hyperbolic Cosine model as an innovative method to identify the appropriateness of feedback. The Hyperbolic Cosine Unfolding Model (HCM; Andrich & Luo, 1993) can be viewed as a probabilistic model for

non-cumulative scales that can be used to identify an ideal point on a continuum that represents a person’s location. It implies a single-peaked response function where a person has a higher probability of endorsing a subset of items, and these items identify the location of a person on the unfolding scale. The probability of endorsement increases when the person’s location gets closer to the item’s location. This feature reflects the basic characteristic of an ideal point process (Coombs, 1964) with the probability of a person endorsing a statement dependent on the distance between the person’s location and the item’s position. The HCM takes the following form (Andrich, 1996; Luo, 2001):

$$P(X_{ij} = k) = \frac{[\cosh(\theta_i - \lambda_j)]^{m-k} \prod_{l=1}^k \cosh(\rho_l)}{\sum_{k=0}^m [\cosh(\theta_i - \lambda_j)]^{m-k} \prod_{l=1}^k \cosh(\rho_l)},$$

when, $k = 0, \prod_{l=1}^k \cosh(\rho_l) \equiv 1$;
 where, in the context of writing,

- $k = 0, \dots, m$, and $m + 1$ is the number of rating categories,
- X_{ij} = observed rating given to student essay i on feedback strategy j ,
- θ_i = location of the student essay i ,
- λ_j = location of the feedback strategy j ,
- ρ_j = threshold/unit parameter for feedback strategy j (these threshold parameters reflect the ZTF for each strategy), and

the underlying unfolding scale for the feedback strategies used by teachers for essays is called the *joint* (J) scale (Coombs, 1964). Each feedback strategy (items) and essay have a unique location on the J scale. The relative distances between feedback strategies and an essay are important and meaningful in the unfolding scaling; therefore, we also construct *individual* (I) scales for each essay by folding the J scale at the ideal point (i.e., HCM location) of each essay on the J scale. The I scale reflects an ordering of strategies based on their relative location for each essay on the unfolding continuum.

4 A General Illustrative Example

An illustrative example is discussed in this section. Let us assume student achievement is represented by different achievement levels as shown in the first scale in Fig. 1. Meanwhile, some feedback strategies focus on different aspects or levels of student achievement. We simulate three possible situations of teachers providing feedback to students: they will not provide a type of feedback strategy, they may provide this type of feedback strategy, and they will provide it to the students. We use a rating scale (0–2) to indicate how likely they would be to provide this type of

feedback at this achievement level: 0 = No, 1 = Possibly, and 2 = Yes. We used the *RateFOLD* program (Luo & Andrich, 2003) to run the HCM analysis.

Teacher responses are shown in Table 2 (Panel A). For student response A, the teacher chooses to use Feedback 1 and 2, possibly to use Feedback 3 and 4, and not to use Feedback 5 and 6. The HCM scale ranges from -3.53 for Feedback 6 to 2.92 for Feedback 1. Table 2 (Panel B) provides information in terms of the distance between feedback strategies and the locations of the answers on the HCM scale. The entries in this table are the absolute values of these distances. Smaller values are highlighted because they are more likely to be endorsed than the others. They indicate the most recommended feedback strategies by the teachers for each essay. For example, the smallest distance for student response B is .16 for Feedback 2 with this being the most recommended feedback strategy. For response C, Feedback 3 is the most recommended one with the smallest distance of .19. Figure 2a shows a HCM map for the feedback strategies. In the HCM map, the ZTF for student response C is highlighted. A useful check on the appropriateness of the HCM model is to fit a polynomial model for the relationship between the location of feedback

Table 2 Illustration based on six feedback strategies for five essays

Panel A: Illustrative ratings						
Feedback strategies	Essays					Feedback location
	A	B	C	D	E	
1	2	1	1	1	0	2.92
2	2	2	2	1	1	1.10
3	1	2	2	2	1	-0.02
4	1	1	2	2	1	-0.56
5	0	1	2	2	1	-1.15
6	0	0	1	1	2	-3.53
Essay location:	2.73	1.26	-0.21	-0.68	-3.10	
Panel B: Absolute values of differences between essays and feedback locations						
Feedback strategies	Essays					Feedback location
	A	B	C	D	E	
1	1.63	1.66	3.13	3.60	6.02	2.92
2	1.63	0.16	1.31	1.78	4.20	1.10
3	2.75	1.28	0.19	0.66	3.08	-0.02
4	3.29	1.82	0.35	0.12	2.54	-0.56
5	3.88	2.41	0.94	0.47	1.95	-1.15
6	6.26	4.79	3.32	2.85	0.43	-3.53
Essay location:	2.73	1.26	-0.21	-0.68	-3.10	

Note: Cell entries are the absolute values of the differences between essay and feedback locations on the unfolding scale. The feedback strategies with smaller distances are highly recommended by the teachers

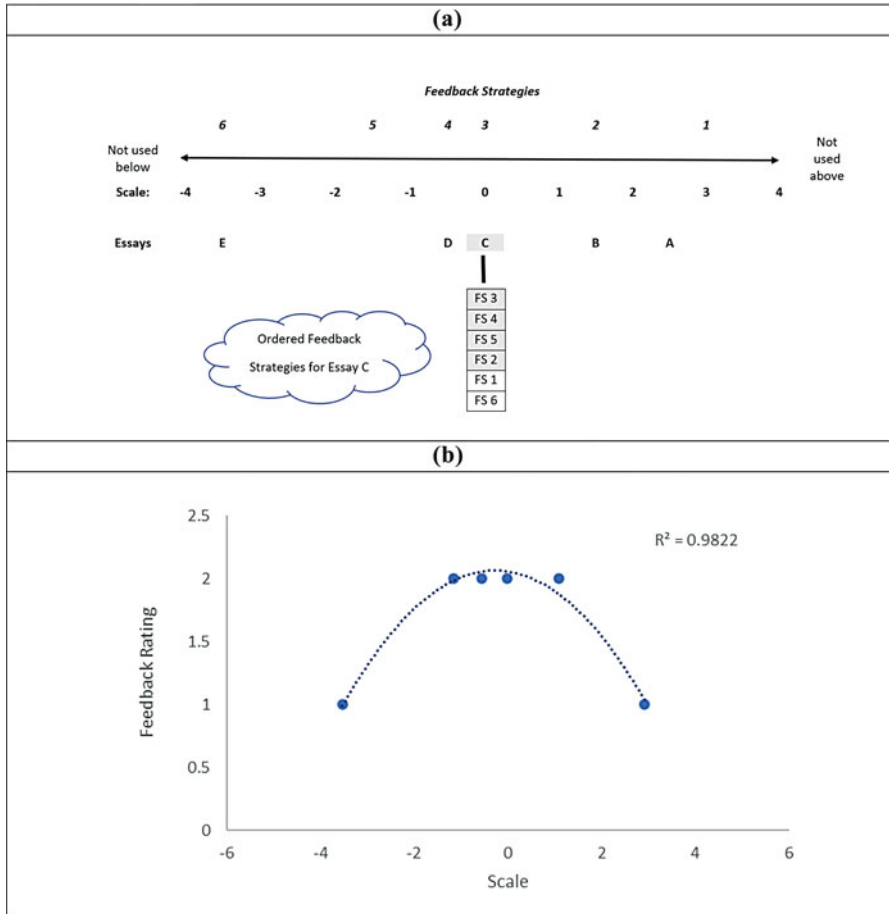


Fig. 2 Illustrative example. (a) Recommend feedback strategies for essay C. (b) Relationship between location for essay C (second degree polynomial)

strategies and the proportion of teacher judgments for each feedback strategy. This is shown in Fig. 2b with a R-square value of .9822 that supports the use of an unfolding process for these feedback data.

5 Application to Writing Assessment

To develop a deeper understanding of the conceptual map and the approach discussed in the previous section, an empirical study that examined feedback in the context of writing assessment is briefly introduced in this section. In this

Table 3 Questionnaire

Short labels	Items (feedback strategies)
<i>A. Organization</i>	
Introduction	1. Feedback should focus on how to create a more effective introduction to the response.
Organization	2. Feedback should focus on how to organize the response more clearly.
Conclusion	3. Feedback should focus on how to create a more effective conclusion to the response.
<i>B. Development</i>	
Relevant evidence	4. Feedback should focus on how to incorporate relevant evidence from the source texts into the response.
Elaborate evidence	5. Feedback should focus on how to elaborate more effectively on the evidence incorporated from the source texts.
Tone	6. Feedback should focus on how to create a tone that is appropriate for the task.
<i>C. Language usage and conventions</i>	
Create sentences	7. Feedback should focus on how to create clear, complete sentences.
Vary sentences	8. Feedback should focus on how to vary sentence structure.
Usage	9. Feedback should focus on usage (e.g., subject-verb agreement, pronoun-antecedent agreement, and using correct forms of homonyms).
Mechanics	10. Feedback should focus on mechanics (e.g., use of internal punctuation, spelling, capitalization, paragraph indentations, etc.)

Notes: Teachers responded to these questions using a 4-point scale indicating how likely they were to provide feedback in each area: 1 = Definitely not, 2 = Probably not, 3 = Probably and 4 = Definitely

empirical study, we conducted both a qualitative design to collecting the data and a quantitative analysis to demonstrate the application of HCM analysis. Essays written by middle school students with different writing proficiencies are used in this study ($N = 20$). A questionnaire was constructed based on a focus group of English teachers' recommendations of the possible feedback strategies for the essays (see Table 3). The ten feedback strategies in the questionnaires focus on three aspects: (1) organization (e.g., feedback should focus on how to create a more effective introduction to the response), (2) development, and (3) language usage (e.g., feedback should focus on how to create clear, complete sentences). Next, middle school ELA teachers ($N = 20$) responded to this questionnaire for identifying feedback strategies for the set of 20 essays. The HCM analysis was done in the *RateFOLD* program after we collect the responses to the questionnaire from the teachers.

It is beyond the scope of this chapter to describe the empirical study results in detail; however, Fig. 3 shows the calibration of essays and feedback on the unfolding continuum. In Table 4, the absolute values of the distances between HCM scale and person ability are showed. Smaller values indicate more highly recommended feedback strategies by the teachers for each essay. Distances less than 2 were highlighted as the ZTF for each essay. These strategies were more likely to

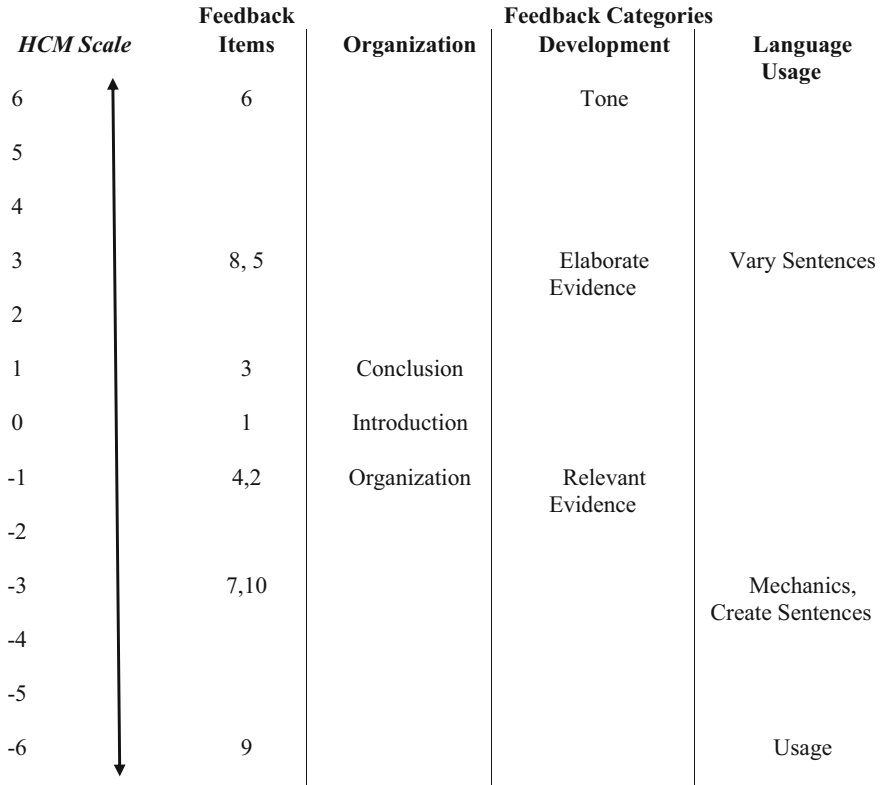


Fig. 3 Map of feedback strategies

be recommended than the others for each essay. The preliminary results from this study suggest that this is a promising approach for identifying ZTF. This program of research uses the HCM to model the recommended feedback strategies, and the next step is to extend the approach to other content areas including mathematics, science, and social studies.

6 Discussion

The study introduces the concept of Zones of Targeted Feedback, and illustrates how to use a Hyperbolic Cosine unfolding model to identify and improve the effectiveness of feedback. The study also briefly presents empirical work in the field of middle school writing assessment. Our illustration indicates that unfolding models can be used as measurement tools to identify the optimal strategies for

Table 4 Results of an empirical study

Feedback items: Scale Essays	9	10	7	2	4	1	3	5	8	6	Essay locations		Chi-squared
											Mean	Variance	
Essay 1	7.05	4.64	4.29	2.61	2.53	1.94	0.64	1.01	1.14	4.05	1.75	3.80	1.81
Essay 2	6.75	4.34	3.99	2.31	2.23	1.64	0.34	1.31	1.44	4.35	1.45	9.99	5.20
Essay 3	6.72	4.31	3.96	2.28	2.20	1.61	0.31	1.34	1.47	4.38	1.42	3.06	2.10
Essay 4	6.51	4.10	3.75	2.07	1.99	1.40	0.10	1.55	1.68	4.59	1.21	9.99	4.54
Essay 5	6.24	3.83	3.48	1.80	1.72	1.13	0.17	1.82	1.95	4.86	0.94	1.80	1.44
Essay 6	6.17	3.76	3.41	1.73	1.65	1.06	0.24	1.89	2.02	4.93	0.87	1.51	.63
Essay 7	6.13	3.72	3.37	1.69	1.61	1.02	0.28	1.93	2.06	4.97	0.83	2.40	2.28
Essay 8	5.54	3.13	2.78	1.10	1.02	0.43	0.87	2.52	2.65	5.56	0.24	0.86	2.24
Essay 9	5.53	3.12	2.77	1.09	1.01	0.42	0.88	2.53	2.66	5.57	0.23	3.92	3.27
Essay 10	5.53	3.12	2.77	1.09	1.01	0.42	0.88	2.53	2.66	5.57	0.23	3.88	1.53
Essay 11	5.45	3.04	2.69	1.01	0.93	0.34	0.96	2.61	2.74	5.65	0.15	5.71	12.36
Essay 12	5.21	2.80	2.45	0.77	0.69	0.10	1.20	2.85	2.98	5.89	-0.09	1.32	1.87
Essay 13	5.11	2.70	2.35	0.67	0.59	0.00	1.30	2.95	3.08	5.99	-0.19	1.66	1.02
Essay 14	4.62	2.21	1.86	0.18	0.10	0.49	1.79	3.44	3.57	6.48	-0.68	2.13	1.69
Essay 15	4.54	2.13	1.78	0.10	0.02	0.57	1.87	3.52	3.65	6.56	-0.76	9.00	18.07 ^a
Essay 16	4.53	2.12	1.77	0.09	0.01	0.58	1.88	3.53	3.66	6.57	-0.77	2.07	2.07
Essay 17	4.32	1.91	1.56	0.12	0.20	0.79	2.09	3.74	3.87	6.78	-0.98	6.76	1.22
Essay 18	4.10	1.69	1.34	0.34	0.42	1.01	2.31	3.96	4.09	7.00	-1.20	1.69	8.50
Essay 19	3.97	1.56	1.21	0.47	0.55	1.14	2.44	4.09	4.22	7.13	-1.33	2.89	9.38
Essay 20	3.82	1.41	1.06	0.62	0.70	1.29	2.59	4.24	4.37	7.28	-1.48	2.25	20.18 ^a

Note: Cell entries are the absolute values of the differences between essay and feedback locations on the HCM scale. The labels for the feedback strategies: 9 = Usage, 10 = Mechanics, 7 = Create Sentences, 2 = Organization, 4 = Relevant Evidence, 1 = Introduction, 3 = Conclusion, 5 = Elaborate Evidence, 8 = Vary Sentences, and 6 = Tone

^a Approximate chi-squared statistics are evaluated with $df = 9$ and $p < .01$

students at different writing proficiency levels, and to define ZTF that may be effective for improving student writing.

Further analyses of ZTF can identify strategies in different content areas. Moreover, examining students' reception of the feedback and exploring empirical evidence of effective learning is essential in the coming future. As researchers generalize from this study to broader contexts, examining appropriate feedback strategies may help fill the gap between formative assessments and what teachers can do in practice to improve student achievement. Research is also needed to explore other unfolding models, such as the nonparametric unfolding IRT model (Post & Snijders, 1993), and the generalized graded unfolding model (Roberts et al., 2000). In summary, the identification and use of zones of targeted feedback (ZTF) offer a promising strategy for moving students forward to a higher level of achievement.

References

- Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology*, *49*, 347–365.
- Andrich, D. (1997). A hyperbolic cosine IRT model for unfolding direct response of persons to items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 399–414). Springer.
- Andrich, D. (1988). The application of an unfolding model of the PIRT type to the measurement of attitude. *Applied Psychological Measurement*, *12*(1), 33–51.
- Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement*, *17*, 253–276.
- Bangert-Drowns, R. L., Kulik, C. L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*(2), 213–238.
- Coombs, C. H. (1964). *A theory of data*. Wiley.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *7*(1), 81–112.
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254.
- Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research*, *47*(2), 211–232.
- Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review*, *1*(4), 279–308.
- Luo, G. (2001). A class of probabilistic unfolding models for polytomous responses. *Journal of Mathematical Psychology*, *45*(2), 224–248.
- Luo, G., & Andrich, D. (2003). *RateFOLD computer program*. Social Measurement Laboratory: School of Education, Murdoch University.
- Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multimedia learning. In *Instructional design for multimedia learning* (pp. 181–195). Waxmann.
- Post, W. J., & Snijders, T. A. (1993). Nonparametric unfolding models for dichotomous data. *Methodika*, *7*(1), 130–156.

- 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.
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.
- Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scriber & E. Souberman, Eds. & Trans.). Harvard University Press
- Willingham, D. B. (1990). Effective feedback on written assignments. *Teaching of Psychology, 17*(1), 10–13.

A Reduced Social Relations Model for Dyad-Constant Dependent Variables



Terrence D. Jorgensen  and K. Jean Forney 

Abstract Dyadic network data occur when each member in a group provides data about each other member in the group (e.g., how much they like each other person). Such data have a complex nesting structure, such that bivariate responses (e.g., Person A's liking of B and vice versa) are dependent upon out-going and in-coming random effects that are correlated within individuals. Dyadic network models for such data include the social relations model for normal data and the p_2 and j_2 models for dichotomous data, but we have seen no application or generalization to accommodate a rarely discussed type of variable from this framework: variables that are constant within a dyad. Dyad-constant variables could include background variables such as whether a dyad is same or opposite sex or how many years two friends have known each other, which require no special modification to use as predictors (Jorgensen et al., Soc Netw 54:26–40, 2018). But they could also be outcomes, such as the difference in a married couple's relationship satisfaction or the similarity in symptoms of a (set of) psychological disorder(s). We explore how such dyad-constant outcomes can be modeled, demonstrating on a data set from a clinic for patients with eating disorders.

Keywords Dyadic data · Social networks · Round-robin designs · Social relations model

T. D. Jorgensen (✉)

University of Amsterdam, Department of Child Development and Education, Amsterdam, The Netherlands

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

K. J. Forney (✉)

Ohio University, Department of Psychology, Athens, OH, USA

e-mail: forney@ohio.edu

1 Introduction

In this paper we present a modified social relations model (SRM) for network-structured outcomes that are constant within a dyad. The SRM is traditionally applied to data gathered from a round-robin design (Warner et al., 1979), wherein each member of a group provides information about each other member of the group (e.g., how much person i likes person j). For a group of size N , this would yield $N^2 - N = N(N - 1)$ unique observations—two observations for each pair (dyad) in the group, depicted by the vector $\mathbf{y}_{\{ij\}}$, where the braces indicate the ordering of members $i \neq j$ is arbitrary. Because each member belongs to multiple dyads, each bivariate dyadic observation is nested in the set of observations in which person i is a member, as well as in the set of observations in which person j is a member.

From a multilevel-modeling perspective (Snijders & Kenny, 1999), the bivariate dyadic (i.e., Level-1) observations $\mathbf{y}_{\{ij\}}$ are cross-classified under both ego and alter (i.e., Level-2) effects. However, round-robin data (also called sociometric data, relational data, interpersonal data, and network data) are more complexly structured than textbook examples of cross-classified data (e.g., students nested with schools and neighborhoods, whose effects are independent of each other). The nature of the network structure is explained by way of introducing the SRM and its extensions, followed by showing how the SRM can be specified to accommodate data that do not vary within a dyad (the focus of this chapter). Results are presented from an empirical clinical-psychology example, which was the motivation behind developing this innovation. The discussion includes comparison with related models and suggestions for future developments.

1.1 The Social Relations Model

The SRM can be depicted as a random-effects model (Nestler, 2016) that decomposes $\mathbf{y}_{\{ij\}}$ into person- and dyad-level components:

$$\mathbf{y}_{\{ij\}} = \begin{bmatrix} y_{ij} \\ y_{ji} \end{bmatrix} = \mu + \begin{bmatrix} E_i + A_j + R_{ij} \\ E_j + A_i + R_{ji} \end{bmatrix}, \quad (1)$$

where μ is the expected value of the observations (e.g., average amount of liking in the group). E_i and A_j are person-level ego (out-going) and alter (in-coming) effects, respectively—for example, E_i would represent how much person i likes others in general, and A_j would represent how much person j is generally liked by others (i.e., likeability). Each R is a dyad-level residual, which contains relationship-specific effects (e.g., how much i uniquely likes j beyond what is expected from their person-level effects) as well as measurement error. More descriptive terms have been used for E_i and A_j , such as actor and partner effects when $\mathbf{y}_{\{ij\}}$ are behavioral interactions (e.g., social mimicry; Salazar Kämpf et al., 2018) or perceiver and target

effects when $\mathbf{y}_{\{ij\}}$ are interpersonal perceptions (e.g., of personality traits; Kenny, 1994).

Each person’s vector of ego and alter effects is assumed bivariate normally distributed:

$$\begin{bmatrix} E_i \\ A_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_{EA} = \begin{bmatrix} \sigma_E^2 & \\ \sigma_{EA} & \sigma_A^2 \end{bmatrix}\right), \tag{2}$$

where σ_E^2 and σ_A^2 are the variances of the random person-level effects, respectively, and $\rho_{EA} = \frac{\sigma_{EA}}{\sigma_E \sigma_A}$ is their correlation, termed *generalized reciprocity* (Kenny, 1994). Following from the liking example, positive generalized reciprocity would be observed when those who have a propensity to (dis)like people are also generally (un)likeable. Negative generalized reciprocity might be observed when measuring helpfulness in a collaborative work situation: those who receive the most help (e.g., because they have lower competence) could be expected to provide the least help to others.

Each dyad’s pair of residuals is also assumed bivariate normally distributed:

$$\begin{bmatrix} R_{ij} \\ R_{ji} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_R = \sigma_R^2 \begin{bmatrix} 1 & \\ \rho_R & 1 \end{bmatrix}\right), \tag{3}$$

where variances σ_R^2 are constrained to equality for indistinguishable dyads (Kenny et al., 2006, ch. 8) and the (residual) correlation ρ_R between relationship effects is called *dyadic reciprocity* (Kenny, 1994). Following from the liking example, positive dyadic reciprocity implies that if person i particularly likes person j (i.e., more than would be expected given person i ’s general propensity for liking and person j ’s general likeability), then person j particularly likes person i , too (i.e., the feeling is mutual). Negative dyadic reciprocity might be observed if person i found person j particularly helpful during collaboration, in which case person j might not have received much help from person i .

A common goal of univariate SRM studies is to calculate the relative contributions of each level of analysis on the overall variability in $\mathbf{y}_{\{ij\}}$ (σ_y^2). This can be expressed as the proportion of variance accounted for by each orthogonal variance component:

$$\sigma_y^2 = \sigma_E^2 + \sigma_A^2 + \sigma_R^2. \tag{4}$$

Note that although E_i and A_i can be correlated *within person i* , the random effects of different persons i and j are assumed independent (and identically distributed). Thus, for a single observation y_{ij} , the ego and alter components are independent. The proportions of variance explained by each component can be used to compare their relative impact on the observed phenomenon. For example, is the degree to which person i likes person j influenced more by person i ’s propensity to like others, by person j ’s likeability, or is it primarily their personal chemistry with each

other? The decomposition of the full covariance matrix for $\mathbf{y}_{\{ij\}}$ (i.e., Σ_y) follows a similar logic:

$$\Sigma_y = \Sigma_{EA} + \Sigma'_{EA} + \Sigma_R, \tag{5}$$

$$\sigma_y^2 \begin{bmatrix} 1 & \\ \rho_{ij} & 1 \end{bmatrix} = \begin{bmatrix} \sigma_E^2 & \sigma_{EA} \\ \sigma_{EA} & \sigma_A^2 \end{bmatrix} + \begin{bmatrix} \sigma_A^2 & \sigma_{EA} \\ \sigma_{EA} & \sigma_E^2 \end{bmatrix} + \sigma_R^2 \begin{bmatrix} 1 & \rho_R \\ \rho_R & 1 \end{bmatrix}. \tag{6}$$

1.2 Extending the SRM with Covariates

Covariates can be added to the SRM, either as explicit predictors of random effects (e.g., Koster & Leckie, 2014; Lüdtke et al., 2013) or as auxiliary correlates (e.g., Brunson et al., 2016), which can alleviate the effects of missing data (Jorgensen et al., 2018). When person-level covariates (\mathbf{x}) are added as predictors of ego and alter effects, the distributional assumption in Eq. 2 applies to Level-2 residuals ε and δ :

$$\begin{bmatrix} E_i \\ A_i \end{bmatrix} = \begin{bmatrix} \sum_{p=1}^P \beta_p x_{i,p} + \varepsilon_i \\ \sum_{p=1}^P \alpha_p x_{i,p} + \delta_i \end{bmatrix}, \tag{7}$$

where P is the number of person-level predictors, β_p is the effect of predictor x_p on ego effects, and α_p is the effect of predictor x_p on alter effects. For example, personality traits (\mathbf{x}) such as openness to experience and extraversion could be used to predict general liking (E) and likability (A), respectively.

Likewise, dyad-level predictors $q = 1, \dots, Q$ can be added to the Level-1 model:

$$\begin{bmatrix} y_{ij} \\ y_{ji} \end{bmatrix} = \beta_0 + \begin{bmatrix} E_i + A_j + \sum_{q=1}^Q \gamma_q w_{ij,q} + \sum_{q=1}^Q \lambda_q w_{ji,q} + R_{ij} \\ E_j + A_i + \sum_{q=1}^Q \gamma_q w_{ji,q} + \sum_{q=1}^Q \lambda_q w_{ij,q} + R_{ji} \end{bmatrix}, \tag{8}$$

where the intercept β_0 is a conditional mean that supplants the role of the grand mean μ in Eq. 1. The E and A terms in Eq. 8 can also incorporate predictors as in Eq. 7. Intrapersonal (γ) and interpersonal (λ) slopes can be distinguished (Nestler, 2016). For example, Salazar Kämpf et al. (2018) reported that person i especially liking person j was associated with person i especially mimicking person j (an intrapersonal effect) in subsequent interaction; however, after a time lag, person i especially mimicking person j was then associated with person j especially liking person i (an interpersonal effect).

Like dyad-level outcomes, dyad-level predictors can differ for each member of the dyad (i.e., $w_{ij} \neq w_{ji}$)—for example, how attractive or agreeable each person thinks the other person is. However, predictors could also be constant within a dyad ($w_{\{ij\}} = w_{ij} = w_{ji}$), which is often (but not necessarily) a function of person-

level variables. For example, a dummy code indicating same- or opposite-sex dyads is a function of the members' sexes. Alternatively, how many months or years the members of a dyad have been acquainted is not a function of their person-level characteristics, but it is nonetheless constant within a dyad. Note that when $w_{ij,q} = w_{ji,q}$, the intrapersonal and interpersonal effects in Eq. 8 cannot be distinguished ($\gamma_q = \lambda_q$), so the predictor $w_{\{ij\},q}$ should only be included once. But even when a predictor W varies within a dyad, intrapersonal (γ) and interpersonal effects (λ) are each constrained to equality across the bivariate observations. These equality constraints hold for indistinguishable dyads, for the same reason there is an equality constraint on the residual variances (i.e., the order within dyad $\{ij\}$ is arbitrary).

2 Reducing the SRM for Dyad-Constant Outcomes

Dependent relational/network variables can also be constant within a dyad ($y_{ij} = y_{ji}$). For example, we might be interested in explaining why friends differ in how much they like each other. No method has been formally defined for accommodating such data, so introducing such a method is the primary goal of this paper. We will begin by focusing only on the basic SRM in Eq. 1, which suffices to discuss the relevant issues. We then consider covariate effects after resolving the issues.

A dyad-constant outcome $y_{\{ij\}}$ is equivalent to equating bivariate observations on the left-hand side of Eq. 1 ($y_{ij} = y_{ji}$), which implies the equality of the summed components on the right-hand side of Eq. 1:

$$E_i + A_j + R_{ij} = E_j + A_i + R_{ji} ; \text{ thus,} \tag{9}$$

$$R_{ij} = R_{ji} \text{ and} \tag{10}$$

$$E_i + A_j = E_j + A_i ; \text{ furthermore,} \tag{11}$$

$$E_i = A_i \text{ and} \tag{12}$$

$$A_j = E_j . \tag{13}$$

Because the person- and dyad-level components of Eq. 9 are independent, this further implies the equivalence of the relationship components in Eq. 10 and of the sum of person components in Eq. 11. Finally, person i 's random effects are independent of person j 's effects, implying the ego and alter effects are equivalent for persons i (Eq. 12) and j (Eq. 13).

Returning to the social-mimicry example (Salazar Kämpf et al., 2018), we might be interested in the degree to which persons i and j differ in how frequently they (un)consciously imitate each other during a conversation. Larger absolute values of this discrepancy ($y_{\{ij\}} = |y_{ij} - y_{ji}|$) could be interpreted as evidence of social dominance within a dyad, whereas smaller absolute values might indicate more equity among conversation partners.

The equivalence of person-level effects ($E_i = A_i$) implies equivalence of their variance components in Eq. 2 ($\sigma_E = \sigma_A$) and a correlation of $\rho_{EA} = 1$. The equivalence of relationship effects (Eq. 10) also implies a correlation of $\rho_R = 1$:

$$\begin{bmatrix} E_i \\ A_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma_E^2 \begin{bmatrix} 1 & \\ & 1 \end{bmatrix}\right) \text{ and } \begin{bmatrix} R_{ij} \\ R_{ji} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma_R^2 \begin{bmatrix} 1 & \\ & 1 \end{bmatrix}\right). \quad (14)$$

This simplifies the distributional assumptions from bivariate normality (Eqs. 2 and 3) to univariate normality:

$$E_i \sim \mathcal{N}(0, \sigma_E) \text{ and } R_{\{ij\}} \sim \mathcal{N}(0, \sigma_R), \quad (15)$$

where the use of E or A subscript is arbitrary given their equivalence, and again the braces around $\{ij\}$ indicate the order is arbitrary. Thus, the dyad-constant outcome $y_{\{ij\}}$ can be expressed as a univariate function of person- and dyad-level effects:

$$y_{\{ij\}} = \mu + E_i + E_j + R_{\{ij\}}. \quad (16)$$

In the social-dominance (discrepancy in mimicry) example, the person-level effect E_i would represent person i 's general tendency to dominate ($y_{ij} - y_{ji} > 0$) or defer ($y_{ij} - y_{ji} < 0$) control of a conversation, and the residual $R_{\{ij\}}$ continues to capture relationship-specific tendencies along with other sources of error.

The reduced SRM in Eq. 16 still contains components from both members of the dyad because the observations $y_{\{ij\}}$ are still nested under (a) all the dyadic observations in which person i is a member, as well as (b) all the dyadic observations in which person j is a member. That is, the observations are still cross-classified, but there is simply no way to distinguish between out-going ego effects and in-coming alter effects. There is simply a single vector of person-level effects, two of which (E_i and E_j) are components of any dyadic observation $y_{\{ij\}}$. Thus, the variance decomposition in Eq. 4 becomes

$$\sigma_y^2 = 2\sigma_E^2 + \sigma_R^2. \quad (17)$$

In practice, the proportion of variance in dyad-constant $y_{\{ij\}}$ attributable to person-level characteristics (e.g., individuals' general tendencies to seize or surrender control of a conversation) should therefore be calculated by doubling the estimated variance of person-level effects: $2\hat{\sigma}_E^2 / (2\hat{\sigma}_E^2 + \hat{\sigma}_R^2)$.

The reduced SRM presented in Eqs. 16 and 15 can be applied to variables $y_{\{ij\}}$ that are identical within a dyad ($y_{ij} = y_{ji}$) or are of equal magnitude but opposite signs ($y_{ij} = -y_{ji}$), which would still be redundant information with the same variance decomposition. Dyad-constant variables could be defined independently, such as the number of things persons i and j found in common when getting acquainted, or $y_{\{ij\}}$ could be a function of person- or dyad-level variables. For example, differences in personality traits (person-level characteristics) would be

dyad-specific but constant in absolute magnitude, as would differences in how much each person likes the other (dyad-level characteristics).

Beyond questions of whether network-level phenomena are driven primarily by person- versus dyad-level characteristics, the reduced SRM provides no information about qualitatively distinct person-level variances ($\sigma_E^2 = \sigma_A^2$), nor does it provide information about generalized or dyadic correlations (both $\rho = 1$ because $E_i = A_i$ and $R_{ij} = R_{ji}$). Thus, the reduced SRM might not garner much interest for substantive applications. When network-structured variables are constant within each dyad, we imagine substantive interest would lie primarily in explaining individual or relationship differences at each level of analysis.

We now extend the reduced SRM to include covariates, which was what motivated its development (see Sect. 3). Given $E_i = A_i$, Eq. 7 reduces to

$$E_i = \sum_{p=1}^P \beta_p x_{i,p} + \varepsilon_i. \tag{18}$$

Substituting the Level-2 model (Eq. 18) into the reduced SRM (Eq. 16) yields an interesting result:

$$\begin{aligned} y_{\{ij\}} &= \beta_0 + E_i + E_j + R_{\{ij\}} \\ &= \beta_0 + \left(\sum_{p=1}^P \beta_p x_{i,p} + \varepsilon_i \right) + \left(\sum_{p=1}^P \beta_p x_{j,p} + \varepsilon_j \right) + R_{\{ij\}} \\ &= \beta_0 + \sum_{p=1}^P \beta_p (x_{i,p} + x_{j,p}) + \varepsilon_i + \varepsilon_j + R_{\{ij\}}, \end{aligned} \tag{19}$$

in that the slope β_p can be multiplied by the sum of person i 's and j 's values on predictor x_p .

Similarly, dyad-level predictors can be added, but when $y_{ij} = y_{ji}$, intra- and interpersonal effects cannot be distinguished ($\gamma = \lambda$), resulting in a similar result as in Eq. 19:

$$\begin{aligned} y_{\{ij\}} &= \beta_0 + E_i + E_j + \sum_{q=1}^Q \gamma_q w_{ij,q} + \sum_{q=1}^Q \gamma_q w_{ji,q} + R_{\{ij\}} \\ &= \beta_0 + E_i + E_j + \sum_{q=1}^Q \gamma_q (w_{ij,q} + w_{ji,q}) + R_{\{ij\}}. \end{aligned} \tag{20}$$

When a predictor is also constant within dyads ($w_{ij,q} = w_{ji,q}$), the slope is effectively a weight for $2 \times w_{\{ij\},q}$. In this case, it would be more intuitive to either (a) divide $w_{\{ij\},q}$ by 2 prior to analysis or (b) only include one ‘‘copy’’ of

$w_{\{ij\},q}$ in Eq. 20, so that $\hat{\gamma}_q$ could be interpreted as expected difference in $y_{\{ij\}}$ per unit-increase in $w_{\{ij\},q}$. In Sect. 3, we use option (b) to estimate the effect of a three-category dyad-constant predictor, represented by two dummy codes.

3 Motivating Example

3.1 Background

Eating disorders are serious psychiatric conditions characterized by overconcern with weight and shape and by problematic behaviors such as fasting, binge-eating, and self-induced vomiting that contribute to increased morbidity and elevated risk of death (van Hoeken and Hoek, 2020). Risk models for eating disorders, such as the Tripartite Influence Model, posit that appearance pressures from peers contribute to the development of eating disorder symptoms (Thompson et al., 1999). Existing research supports the importance of both socialization (i.e., contagion) of some disordered eating behaviors (i.e., binge-eating like behaviors; Zalta and Keel, 2006), and selection effects, that is, individuals forming friendships with those who share similar attitudes (Rayner et al., 2013). If peer attitudes and behaviors increase risk for eating-disorder attitudes and behaviors, then peers might also be able to help mitigate risk or clinically significant eating-disorder symptoms. Indeed, changes in perceptions of peer norms predict decreases in disordered eating attitudes in prevention settings (Cruwys et al., 2015).

While peers are prominent in sociocultural models of eating disorder risk, less is known about the role of peers in the maintenance or treatment of clinically diagnosed eating disorders. Outpatient treatment models rely on one-to-one interactions between patients and clinicians. Augmenting this standard care with mentorship from a recovered peer improves some treatment outcomes (Ranzenhofer et al., 2020). More intensive treatments, such as partial hospitalization programs, occur in group settings for significant periods of time (e.g., 30 hours/week). By the nature of the group setting, peers with eating disorders are an integral part of treatment in partial hospitalization programs. Preliminary work suggests that patients who develop quality friendships in treatment have greater motivation to change (Malmendier-Muehlschlegel et al., 2016). However, engagement in relationships developed during treatment is associated (after discharge) with both positive and negative outcomes, depending on the types of interactions that take place (Saffran et al., 2016).

Taken together, the literature suggests that friendships—and the disordered eating attitudes and behaviors of those friends—may play a role in the maintenance of eating disorder behaviors and thus may also play a role in facilitating positive treatment outcomes. Existing literature has been limited by reliance on perceptions of peers, but social network data can overcome these limitations by “objectively” measuring peer eating-disorder symptoms (Jorgensen et al., 2018). The current

study represents the first step in understanding how friendships formed in treatment may contribute improved treatment outcomes via the modeling of socialization and recovery-oriented attitudes and behaviors. We sought to understand whether similarity in the severity of eating disorder symptoms was associated with reported friendships. We hypothesized that eating disorder symptoms would be more similar among friends than non-friends.

3.2 Method

Participants and Procedure Participants were recruited from a nonprofit partial hospitalization program for eating disorders in the midwestern United States. Patients were enrolled in 30 hours of evidence-based treatment per week, with most treatment occurring in a group setting. As part of routine care, patients completed assessments on a weekly basis. For the current study, participants were provided a letter from the second author inviting the participant to participate in a study of social influences on eating-disorder treatment outcomes. A graduate research assistant was available to answer any questions that patients had about the research study. After having questions answered, participants provided written informed consent. All study procedures were approved by the local Institutional Review Board. Once enrolled, participants were asked to complete a weekly assessment of eating disorder symptoms and a social network assessment. In addition, participants were asked to provide permission to access their medical record to extract diagnosis, demographic information, and weekly assessments of depressive and anxiety symptoms. Data collection occurred over an 8-week period. Over the 8-week period, 18 individuals were invited to participate, 13 individuals provided informed consent, and 12 individuals provided data on their eating disorder symptoms and social network. Because patients both began and ended treatment during the course of the study, participation by week ranged from 3 to 7. The current study focuses on the three consecutive weeks with highest absolute participation. Among those who provided data, mean age was 27.25 years ($SD = 14.59$). All participants identified as non-Hispanic, White females. The modal eating-disorder diagnosis was “Other Specified Feeding or Eating Disorder,” and all patients had a comorbid mood or anxiety disorder.

Measures Eating-disorder symptoms were measured using the Eating Disorder Examination Questionnaire Short Form (Gideon et al., 2016), an adapted version of the well-established Eating Disorder Examination Questionnaire (Fairburn and Beglin, 1994). Twelve questions assess eating-disorder attitudes and behaviors over the previous 7 days; possible scores ranged 0–36. Scale reliability ranged from $\alpha = 0.89$ – 0.95 across weeks.

Participants were provided a roster of all patients in the partial hospitalization program. They were asked to identify “your friends, that is, the group members you hang around with the most or are closest to. You are welcome to list as many friends

as appropriate.” They were also given the option to identify group members whom they looked to as a role model for recovery.

3.3 Estimating a Reduced SRM

Our outcome of interest was the absolute difference in eating-disorder symptoms between patients i and j (i.e., $y_{\{ij\}}$ was a function of person-level characteristics). Higher values indicated greater dissimilarity in a dyad, and values closer to zero indicated greater similarity. Our focal predictor was whether patients i and j considered each other friends, which is a dyad-level variable that can vary within dyads. Using mutual nonfriendship as the reference category, we used a dummy code to indicate dyads with reciprocated friendship ($w_{\{ij\},2}$). Because the outcome did not vary with dyads, we could not distinguish between asymmetry in one direction (ij) or the other (ji), so a single dummy code ($w_{\{ij\},1}$) represented nonreciprocal, asymmetric friendship. Thus, in our fitted model:

$$y_{\{ij\}} = |y_i - y_j| = \beta_0 + \gamma_1 w_{\{ij\},1} + \gamma_2 w_{\{ij\},2} + E_i + E_j + R_{\{ij\}}, \quad (21)$$

β_0 represents the average dissimilarity in eating disorder symptoms among mutual nonfriends. The slopes for dummy codes in Eq. 21 represent how different from nonfriends the average (dis)similarity was among asymmetric (γ_1) and mutual (γ_2) friends. Our fitted model does not include person-level effects.

Given the documented limitations of two-step estimation approaches (Nestler, 2016; Nestler et al., 2020; Lüdtke et al., 2018), we only considered options to estimate the whole model simultaneously. Although maximum likelihood estimation (MLE) is available for round-robin data in the R package `srm` (Nestler et al., 2019), the software is not set up to accommodate dyad-constant outcome variables (Nestler et al., 2020). Instead, we used Markov chain Monte Carlo (MCMC) estimation (Hoff, 2005; Lüdtke et al., 2013; Jorgensen et al., 2018) using the general Bayesian modeling software Stan (Carpenter et al., 2017) via the R package `rstan`. Stan uses a modified Hamiltonian Monte Carlo (HMC) algorithm called the No-U-Turn Sampler (NUTS) that simultaneously samples the entire vector of estimates from the parameter space, as opposed to iterating one parameter at a time like Gibbs sampling.

The unknown quantities (parameters) in Eq. 21 include β_0 , the fixed effects (γ_1 and γ_2), the vector of person-level random effects (E), and the variance components (σ_E^2 and σ_R^2). In our Stan program (available online¹ and in the Appendix), we specified a standard-normal prior distribution for random effects:

¹ Data and software scripts are available on the Open Science Framework (OSF): <https://osf.io/j53n8/>. No person-level variables are provided, and person-level IDs are randomized within each week to preserve anonymity (i.e., IDs in Week 6 do not correspond to IDs in Weeks 7 or 8).

$$E^* \sim \mathcal{N}(\mu = 0, \sigma = 1). \quad (22)$$

Priors for other parameters were selected based on descriptive statistics to be minimally informative without placing undue weight on values far outside the range of data (Smid et al., 2020). We specified half-normal priors (i.e., normal distributions truncated below 0) for *SDs* rather than variances (σ_E and σ_R):

$$\sigma_{(E \text{ or } R)} \sim \text{half-}\mathcal{N}\left(\frac{h}{5}, \frac{h}{5}\right), \quad (23)$$

where $h = \frac{\max(y_{ij}) - \min(y_{ij})}{2}$ (i.e., half the empirical range of the outcome). The intercept and slopes were also specified to include the range of plausible values without being strongly informative:

$$\beta_0 \sim \mathcal{N}(\text{median}(y_{ij}), h) \quad (24)$$

$$\gamma_{(1 \text{ or } 2)} \sim \mathcal{N}(0, h) \quad (25)$$

To calculate each dyad's expected values \hat{y}_{ij} each time parameters were sampled from the posterior (indexed below with superscript m), we added scaled random effects ($E = E^* \times \sigma_E$) to the intercept and fixed effects (i.e., Eq. 21 but with the residual R_{ij} omitted).

$$\hat{y}_{ij}^m = \beta_0^m + \gamma_1^m w_{\{ij\},1} + \gamma_2^m w_{\{ij\},2} + (E_i^{*,m} + E_j^{*,m})\sigma_E^m, \quad (26)$$

The likelihood was thus specified as:

$$y_{ij} \sim \mathcal{N}(\hat{y}_{ij}, \sigma_R). \quad (27)$$

After 250 burn-in iterations on each of 4 Markov chains, we saved 250 samples from each chain's estimated posterior distribution. Convergence was assessed visually by verifying proper mixing in traceplots, as well as numerically using the potential scale-reduction factor ($\hat{R} < 1.05$; Gelman and Rubin, 1992) and effective sample size ($N_{\text{eff}} > 100$; Vats et al., 2019), both of which are reported with results in Table 1. The combined 1000 samples from the posterior were used to calculate point (posterior mean) and *SE* (posterior *SD*) estimates, as well as empirical 95% credible intervals (CIs) for each (function of) parameter(s). The difference between slopes ($\gamma_2 - \gamma_1$) was calculated at each iteration to capture the mean difference between reciprocal and asymmetric friendships. The 95% CIs were used to infer whether differences between groups were (non)zero.

3.4 Results and Discussion

Variability of friendship nominations confirmed that friendships form in partial hospitalized programs, although friendship nominations were not always reciprocated. Estimated group mean differences in dissimilarity of eating-disorder symptoms are presented in Table 1. Recall that symptoms were measured on a 0–36 scale, so absolute differences between subjects could be as large as 36, although they tended to be smaller because all patients had relatively higher symptom-scores than would be expected in the general population. The top row indicates that due to autocorrelation, the 1000 samples from the posterior have an effective sample size of 421 independent samples from the posterior, which is more than sufficient to minimize Monte Carlo sampling error. The small $\widehat{R} = 1.01$ shows no evidence of convergence problems, and the traceplots (not shown here, but available using the R script on OSF) show evidence of good mixing across chains. Similar convergence diagnostics were found across parameters (all $\widehat{R} \leq 1.01$), and all other effective sample sizes in Table 1 exceed 500.

Again focusing on the top row of Table 1, the average dissimilarity between patients i and j was 2.46 points higher ($SE = 2.42$) among asymmetric friends than among nonfriends. The corresponding CI indicates that given the observed data, there is a 95% posterior probability that the true mean difference is between -2.12 and 7.27 , which is a very wide margin of error. Likewise, the second row shows 1.92 points higher average dissimilarity ($SE = 2.09$) among mutual friends than among nonfriends, also with a large margin of error: 95% CI $[-2.27, 5.73]$. These results contradict our hypothesis that friends would manifest more similar symptoms, but because both CIs include 0, we cannot reject the H_0 that there is simply no effect of friendship on (dis)similarity. The third row compares asymmetric to mutual friends

Table 1 MCMC summaries from fitting reduced SRM to cross-sectional samples from the 3 weeks with the largest sample sizes (21 dyads). Pairwise comparisons are made between groups of dyads that indicated no friendship (0), asymmetric friendship (1), or reciprocated friendship (2). EAP = expected a posteriori (posterior mean), SD = posterior standard deviation, CI = credible interval calculated from posterior percentiles, N_{eff} = effective number of posterior samples (given autocorrelation), \widehat{R} = potential scale-reduction factor

Week	Groups	EAP	SD	95% CI	N_{eff}	\widehat{R}
6	1 vs. 0	2.46	2.42	$[-2.12, 7.27]$	421.06	1.01
	2 vs. 0	1.92	2.09	$[-2.27, 5.73]$	501.27	1.01
	2 vs. 1	-0.54	2.18	$[-5.14, 3.95]$	768.25	1.00
7	1 vs. 0	3.10	5.53	$[-7.99, 13.48]$	1176.77	1.00
	2 vs. 0	-1.00	3.60	$[-8.10, 6.16]$	961.92	1.00
	2 vs. 1	-4.10	6.54	$[-16.35, 9.10]$	1304.70	1.00
8	1 vs. 0	-3.35	5.23	$[-13.26, 7.24]$	722.00	1.00
	2 vs. 0	-3.00	3.79	$[-10.73, 4.31]$	602.52	1.01
	2 vs. 1	0.35	5.81	$[-11.42, 11.17]$	1121.66	1.00

in Week 6, and the remaining rows of Table 1 show the estimated mean differences between types of dyad for Weeks 7 and 8.

Across weeks, no clear pattern emerged in eating disorder symptoms; friends were not consistently more or less similar in regards to their eating disorder symptoms than nonfriends. Whereas Week-6 results descriptively indicate that asymmetric (1 vs. 0) and mutual friendship (2 vs. 0) led to greater dissimilarity, Week-8 results showed that asymmetric and mutual friendship led to greater similarity; Week-7 results were mixed. However, all 95% CIs revealed quite a lack of precision, so the none of the differences could be distinguished statistically from zero.

Friendship formation may be more strongly related to other traits, such as personality factors (Forney et al., 2019) or other aspects of psychopathology, rather than to eating disorder symptoms. Indeed, a more consistent pattern was observed such that friends tended to be more similar on depressive symptoms than nonfriends (see the file “SupplementalResults.pdf” on OSF). Additionally, prior work supports that the make-up of a therapy group (i.e., between-group effects) has a moderate effect on treatment outcomes (Kivlighan et al., 2020). Thus, the lack of consistent findings from week to week may reflect changes in the group make-up as patients entered or left treatment or got to know one another better. Future work may wish to examine closeness as a moderator of any similarity effects. Collecting and combining data from multiple, independent partial hospitalization programs will allow for a better understanding of factors that influence whether or how eating disorder symptoms are related to friendship formation and the converse: whether or how friendship may be related to improvements in eating disorder symptoms.

4 General Discussion

This chapter presents a reduced SRM with covariates designed to enable modeling of dyad-level outcomes that do not vary within a dyad. The real-data application showed that the model is estimable with real data, and converges quickly on a solution, even with relatively little data. However, the low sample size each week did not provide much power to detect any of the estimated effects. Future research should verify the practical feasibility of this model via Monte Carlo simulation studies.

Although it is difficult to anticipate the demand for this model development, it is noteworthy that similar network models for binary outcomes—the p_2 (Van Duijn et al., 2004; Zijlstra et al., 2006) and j_2 (Zijlstra, 2017) models—have also been adapted for symmetric (dyad-constant) outcomes (Blanken et al., 2021), as implemented in the `b2ML()` function of the R package `dyads` (Zijlstra, 2021). Future research might compare this implementation to the reduced SRM presented here, but with a probit link (Koster & Aven, 2018) to accommodate a binary outcome.

Appendix

Annotated Stan (Carpenter et al., 2017) syntax for the SRM fitted to the motivating-example data in Sect. 3.2 is provided below. The *.stan file can be found on the OSF, along with the data and an R script to fit the model using the R package `rstan`: <https://osf.io/j53n8/>

```

data {
  // sample sizes
  int<lower=0> Nd;           // number of dyads (Level 1)
  int<lower=0> Np;           // number of persons (Level 2, cross-classified)
  // observed data
  vector[Nd] Y;             // observed round-robin outcome
  vector[Nd] one;           // dummy codes: one-way friend nomination
  vector[Nd] both;          // reciprocal friend nomination
  // ID variables above Level 1
  int IDp[Nd, 2];           // person-level IDs (cross-classified)
}
transformed data {
  // save limits for default priors
  vector[2] yLimits;
  real halfRange;

  // calculate observed limits
  yLimits[1] = min(Y);
  yLimits[2] = max(Y);
  halfRange = (yLimits[2] - yLimits[1]) / 2;
}
parameters {
  // means and SDs
  vector[3] BETA;           // intercept + 2 slopes
  real<lower=0> s_d;         // dyad-level residual SD
  real<lower=0> s_p;         // person-level random-effect SD

  vector[Np] e_p;           // vector of person-level random effects (unit scale)
}
transformed parameters {
  vector[Nd] Yhat;          // expected values, given random effects
  for (n in 1:Nd) {
    Yhat[n] = BETA[1] + BETA[2]*one[n] + BETA[3]*both[n] +
      s_p*e_p[ IDp[n,1] ] + s_p*e_p[ IDp[n,2] ];
  }
}
model {
  // priors for means/slopes and SDs, based on empirical ranges:
  BETA[1] ~ normal(yLimits[1] + halfRange, halfRange);
  BETA[2] ~ normal(0, halfRange);
  BETA[3] ~ normal(0, halfRange);
  s_d ~ normal(halfRange / 5, halfRange / 5) T[0, ]; // residual SD
  s_p ~ normal(halfRange / 5, halfRange / 5) T[0, ]; // person-level SD
  // random effects (sample on unit scale)
  e_p ~ std_normal();

  // likelihood
  Y ~ normal(Yhat, s_d);
}
generated quantities{
  // mean difference between groups with dummy codes
  real recip;
  recip = BETA[3] - BETA[2];
}

```

References

- Blanken, T. F., Tanis, C. C., Nauta, F. H., Dablander, F., Zijlstra, B. J., Bouten, R. R., Oostvogel, Q. H., Boersma, M. J., van der Steenhoven, M. V., van Harreveld, F., de Wit, S., & Borsboom, D. (2021). Promoting physical distancing during COVID-19: A systematic approach to compare behavioral interventions. *Scientific Reports*, *11*(19463), 1–8.
- Brunson, J. A., Øverup, C. S., & Mehta, P. D. (2016). A social relations examination of neuroticism and emotional support. *Journal of Research in Personality*, *63*, 67–71.
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32.
- Cruwys, T., Haslam, S. A., Fox, N. E., & McMahon, H. (2015). “that’s not what we do”: Evidence that normative change is a mechanism of action in group interventions. *Behaviour Research and Therapy*, *65*, 11–17.
- Fairburn, C. G., & Beglin, S. J. (1994). Assessment of eating disorders: Interview or self-report questionnaire? *International Journal of Eating Disorders*, *16*(4), 363–370.
- Forney, K. J., Schwendler, T., & Ward, R. M. (2019). Examining similarities in eating pathology, negative affect, and perfectionism among peers: A social network analysis. *Appetite*, *137*, 236–243.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.
- Gideon, N., Hawkes, N., Mond, J., Saunders, R., Tchanturia, K., & Serpell, L. (2016). Development and psychometric validation of the EDE-QS, a 12 item short form of the eating disorder examination questionnaire (EDE-Q). *PLoS One*, *11*(5), e0152744.
- Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, *100*(469), 286–295.
- Jorgensen, T. D., Forney, K. J., Hall, J. A., & Giles, S. M. (2018). Using modern methods for missing data analysis with the social relations model: A bridge to social network analysis. *Social Networks*, *54*, 26–40.
- Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York: Guilford.
- Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. Guilford, New York.
- Kivlighan, III, D. M., Aloe, A. M., Adams, M. C., Garrison, Y. L., Obrecht, A., Ho, Y. C. S., Kim, J. Y. C., Hooley, I. W., Chan, L., & Deng, K. (2020). Does the group in group psychotherapy matter? A meta-analysis of the intraclass correlation coefficient in group treatment research. *Journal of Consulting and Clinical Psychology*, *88*(4), 322–337.
- Koster, J., & Aven, B. (2018). The effects of individual status and group performance on network ties among teammates in the National Basketball Association. *PLoS One*, *13*(4), e0196013.
- Koster, J. M., & Leckie, G. (2014). Food sharing networks in lowland Nicaragua: An application of the social relations model to count data. *Social Networks*, *38*, 100–110.
- Lüdtke, O., Robitzsch, A., Kenny, D. A., & Trautwein, U. (2013). A general and flexible approach to estimating the social relations model using Bayesian methods. *Psychological Methods*, *18*(1), 101–119.
- Lüdtke, O., Robitzsch, A., & Trautwein, U. (2018). Integrating covariates into social relations models: A plausible values approach for handling measurement error in perceiver and target effects. *Multivariate Behavioral Research*, *53*(1), 102–124.
- Malmendier-Muehlschlegel, A., Rosewall, J. K., Smith, J. G., Hugo, P., & Lask, B. (2016). Quality of friendships and motivation to change in adolescents with Anorexia Nervosa. *Eating Behaviors*, *22*, 170–174.
- Nestler, S. (2016). Restricted maximum likelihood estimation for parameters of the social relations model. *Psychometrika*, *81*(4), 1098–1117.
- Nestler, S., Lüdtke, O., & Robitzsch, A. (2020). Maximum likelihood estimation of a social relations structural equation model. *Psychometrika*, *85*(4), 870–889.

- Nestler, S., Robitzsch, A., & Luedtke, O. (2019). SRM: Structural equation modeling for the social relations model. R package version 0.3-6.
- Ranzenhofer, L. M., Wilhelmly, M., Hochschild, A., Sanzone, K., Walsh, B. T., & Attia, E. (2020). Peer mentorship as an adjunct intervention for the treatment of eating disorders: A pilot randomized trial. *International Journal of Eating Disorders*, *53*(5), 767–779.
- Rayner, K. E., Schniering, C. A., Rapee, R. M., Taylor, A., & Hutchinson, D. M. (2013). Adolescent girls' friendship networks, body dissatisfaction, and disordered eating: Examining selection and socialization processes. *Journal of Abnormal Psychology*, *122*(1), 93–104.
- Saffran, K., Fitzsimmons-Craft, E. E., Kass, A. E., Wilfley, D. E., Taylor, C. B., & Trockel, M. (2016). Facebook usage among those who have received treatment for an eating disorder in a group setting. *International Journal of Eating Disorders*, *49*(8), 764–777.
- Salazar Kämpf, M., Liebermann, H., Kerschreiter, R., Krause, S., Nestler, S., & Schmukle, S. C. (2018). Disentangling the sources of mimicry: Social relations analyses of the link between mimicry and liking. *Psychological Science*, *29*(1), 131–138.
- Smid, S. C., McNeish, D., Miočević, M., & van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling*, *27*(1), 131–161.
- Snijders, T. A., & Kenny, D. A. (1999). The social relations model for family data: A multilevel approach. *Personal Relationships*, *6*(4), 471–486.
- Thompson, J. K., Heinberg, L. J., Altabe, M., & Tantleff-Dunn, S. (1999). *Exacting beauty: Theory, assessment, and treatment of body image disturbance*. Washington: American Psychological Association.
- Van Duijn, M. A., Snijders, T. A., & Zijlstra, B. J. (2004). p_2 : A random effects model with covariates for directed graphs. *Statistica Neerlandica*, *58*(2), 234–254.
- van Hoeken, D., & Hoek, H. W. (2020). Review of the burden of eating disorders: Mortality, disability, costs, quality of life, and family burden. *Current Opinion in Psychiatry*, *33*(6), 521–527.
- Vats, D., Flegal, J. M., & Jones, G. L. (2019). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika*, *106*(2), 321–337.
- Warner, R. M., Kenny, D. A., & Stoto, M. (1979). A new round robin analysis of variance for social interaction data. *Journal of Personality and Social Psychology*, *37*(10), 1742–1757.
- Zalta, A. K., & Keel, P. K. (2006). Peer influence on bulimic symptoms in college students. *Journal of Abnormal Psychology*, *115*(1), 185–189.
- Zijlstra, B. J. (2017). Regression of directed graphs on independent effects for density and reciprocity. *The Journal of Mathematical Sociology*, *41*(4), 185–192.
- Zijlstra, B. J. (2021). *dya*_{ds}: Dyadic network analysis. R package version 1.1.4.
- Zijlstra, B. J., Van Duijn, M. A., & Snijders, T. A. (2006). The multilevel p_2 model. *Methodology*, *2*(1), 42–47.

Quality Assurance in Digital-First Assessments



Manqian Liao, Yigal Attali, Alina A. von Davier, and J. R. Lockwood

Abstract Computational psychometrics, a blend of theory-driven psychometrics and data-driven algorithms, provides the theoretical underpinnings for the design and analysis of the new generation of high-stakes, digital-first assessments that can be taken anytime and anywhere in the world, and their scores impact test takers' lives. The unprecedented flexibility, complexity, and high-stakes nature of these digital-first assessments pose enormous quality assurance challenges. In order to ensure these assessments meet both “the contest and the measurement” requirements of high-stakes tests, it is necessary to conduct continuous pattern monitoring and to be able to promptly react when needed. In this paper, we illustrate the development of a quality assurance system for a high-stakes and digital-first assessment. To build the system, educational data from continuous administrations of the assessments are mined, modeled and monitored. In particular, five categories of statistics are monitored to assure the quality of the assessment, including scores, test taker profiles, repeaters, item analysis and item exposure. Various control charts and models were applied to detect and flag the abnormal changes in the assessment statistics. The monitoring results and alerts were communicated with the stakeholders via an interactive dashboard. The paper concludes with a discussion on how the automatic quality assurance system is combined with the human review process in real-world application.

Keywords Quality assurance · Digital-first assessment · High-stakes assessment

1 Introduction

Digital-first assessments are assessments that are delivered online and can be taken anywhere and anytime. The flexibility of digital-first assessments results in an optimized test taker experience (Burstein et al., 2021). The development and

M. Liao (✉) · Y. Attali · A. A. von Davier · J. R. Lockwood
Duolingo, Inc., Pittsburgh, PA, USA
e-mail: mancy@duolingo.com; yigal@duolingo.com; avondavier@duolingo.com;
jr@duolingo.com

delivery of digital-first assessments are dependent on a wide range of digital tools, including automatic systems for test development, scoring, and test delivery. In contrast to traditional large-scale assessments that are based on in-person administration to large groups of test takers in fixed locations, digital-first assessments are administered continuously to individual test takers. The advantages of digital-first assessments have manifested themselves during the pandemic, when traditional group assessments in brick-and-mortar test centers became impractical. When digital-first assessments are used for high-stakes purposes (e.g., college admission, employment), they have significant impact on test takers' lives. Therefore, a rigorous quality assurance (QA) system is required to ensure the integrity and validity of the test scores.

QA refers to a systematic process for maintaining the high quality of the test and assessment scores and to prevent errors at all stages of the test, including test design, item design and development, test scoring, test analysis and score reporting (International Test Commission, 2014). QA is a key step to ensure that the digital-first high-stakes assessments meet both “the contest and the measurement” requirements of high-stakes tests (Holland, 1994), where the “contest” refers to the expectation that the test gives everyone a fair chance, and the “measurement” refers to the requirement that the test is accurate and valid.

Since digital-first assessments differ from traditional assessments in many aspects (e.g., administration frequency, item bank size), it is challenging but necessary to develop a QA system that is tailored for digital-first assessments. Such a QA system must serve two purposes. On one hand, it must keep track of statistics to evaluate the health of the test. On the other hand, it must raise alerts when there are irregularities in the statistics. To build such a system, at least four design questions need to be addressed: (1) What statistics should be tracked in the quality assurance system? (2) How can the statistics be continuously updated and monitored? (3) What criteria should be used to identify irregularities in the statistics? and (4) How can the alerts be communicated and addressed? This paper is aimed at documenting the development process of a QA system called Analytics for Quality Assurance in Assessment (AQuAA) for a real-world, digital-first high-stakes language assessment. Given the data-rich nature of digital-first assessments, computational psychometrics (von Davier, 2015, 2017; von Davier et al., 2021), a blend of theory-driven psychometrics and data-driven algorithms, is leveraged to develop AQuAA. Considerations for each of the four design questions above are elaborated upon. While each assessment is unique, it is hoped that this paper can serve as a framework to help future practitioners to develop QA systems for their own digital-first assessments.

1.1 Theoretical Framework

Quality assurance plays an important role in maintaining test score validity. Allalouf (2007) indicated that mistakes that jeopardize assessment score validity could occur

at all stages of assessment development and administration, and that mistakes could accumulate since many stages are contingent on previous stages. Therefore, quality control guidelines and step-by-step procedures (Allalouf, 2007; Allalouf et al., 2017; International Test Commission, 2014) have been developed to help test developers identify possible mistakes as well as their causes, thereby helping them to identify solutions to fix the mistakes and prevent them from happening again.

Quality control procedures were primarily designed for traditional large-scale assessments that are administered on only a few test dates and have large test volumes in each administration (e.g., Allalouf, 2007), with Allalouf et al. (2017) being an exception. Allalouf et al. (2017) recommended a quality control procedure for continuous-mode tests (i.e., tests that are administered to small groups of test takers on many test dates) which share some similarities with digital-first assessments. Moreover, Allalouf et al. (2017) have demonstrated an automated quality control system for continuous-mode tests and the system consists of both an automatic component and a human review component. These two parts also apply to the quality assurance of digital-first assessments. In the automatic part, a number of steps that need to be conducted recurrently and which can be implemented programmatically are packed into an automatic procedure with the use of digital tools. Steps in such an automatic procedure may include fetching the data from the database, conducting a variety of quality control analyses (see Lee & von Davier, 2013 for a review of quality control methods) and generating statistical reports. In the human review part, human experts are trained to review the statistical reports generated from the automatic procedure in order to identify the sources for the potential irregularities or outliers, and determine whether or what actions need to be taken to handle these irregularities.

The foundation of an automated quality assurance procedure consists of a wide range of data mining and data visualization techniques. In the realm of quality assurance, the data mining and data visualization techniques serve two major purposes: First, to describe the trends and seasonal patterns of the assessment statistics; Second, to detect abrupt changes in the relevant assessment statistics. Lee and von Davier (2013) have summarized a number of statistical methods and data visualization techniques for score quality assurance purposes. Various time series techniques can be chosen to describe trends or seasonal patterns, which include linear ANOVA models (Haberman et al., 2008), regression with autoregressive moving-average (Li et al., 2009), harmonic regressions (Lee & Haberman, 2013) and dynamic linear models (Wanjohi et al., 2013). The Shewhart chart is a useful data visualization tool for continuous test score characteristics (Schafer et al., 2011). In terms of detecting abrupt changes in the assessment statistics, some model-based approaches have been applied to mine the data and identify abrupt changes in score time series, such as change-point models and hidden Markov model (Lee & von Davier, 2013). A data visualization technique for detecting abrupt changes is cumulative sum (CUSUM) charts (Page, 1954).

The products of the automated QA procedure may include summary tables of the statistics, graphs and statistical testing results (Allalouf et al., 2017). These statistical products could be organized into different formats, such as reports

(Allalouf et al., 2017) and dashboards (Mohadjer & Edwards, 2018). Since the products of the automated quality assurance procedure serve as the starting point of the human review process (Allalouf et al., 2017), the choice of organizing format should be determined by the ease of communication to the targeted stakeholders.

1.2 The Context of AQuAA

The Duolingo English Test is a high-stakes computerized adaptive test that is designed to be accessible anywhere and anytime (Settles et al., 2020). Thus, it falls under the category of continuous-mode assessments (Allalouf et al., 2017). The Duolingo English Test is an adaptive test, with a large item bank that has been designed by subject matter experts (SMEs) and produced automatically by computer. The items are reviewed by panels of SMEs to ensure quality and cultural fit. The items are scored automatically, and the scoring methods are reviewed periodically by SMEs. Each individual test is proctored remotely using an asynchronous system that involves both AI-based tools and human proctors. Certified test takers are expected to get their test scores within 48 hours after they take the test. The role of AQuAA is to ensure that all test scores are valid and of high quality.

2 What Statistics Should Be Tracked in the Quality Assurance System?

To determine what statistics should be tracked in AQuAA, we need to identify statistics that indicate test quality and score validity of the assessments. As noted by Kane (2013), the interpretation and use of test scores should be validated from various aspects, such as scoring and generalization. Accordingly, AQuAA needs to monitor multiple categories of statistics. In the context of the Duolingo English Test, we determine that the following five categories of statistics cover most aspects of test quality and are feasible to be monitored continuously: scores, test taker profile, repeaters, item analysis, and item exposure.

2.1 Scores

Test scores are directly used by stakeholders (e.g., test takers, institutions), thus summary statistics of the scores, including overall scores, sub-scores, and item type scores, are tracked in AQuAA. Score-related statistics include the location and spread of scores, inter-correlations among scores, bivariate or multivariate outliers,

internal consistency measures, standard error of measurement (SEM) and validity coefficients (e.g., correlation with self-reported external measures).

2.2 Test Taker Profile

The composition of the test taker population is tracked over time, as it could be used to explain the variability in test scores to some extent. Specifically, the (percentage) volume of test takers in important population categories, such as country, native language, gender, age, and intent in taking the test, are tracked. In addition, many of the score statistics are tracked across major test-taker groups.

2.3 Repeaters

In AQuAA, repeaters are defined as those test takers who take the test more than once within a 30-day window. The prevalence, composition, and performance of the repeaters are tracked. The composition of the repeater population is defined with respect to the same test taker profile categories discussed above; the performance of the repeater population is tracked with many of the same test score statistics identified above. Statistics that are specific to the “repeaters” category include location and spread of the first and second test scores, score difference across test attempts, test-retest reliability and SEM.

2.4 Item Analysis

Ensuring that items are of high quality and that the item quality is stable over time are the prerequisites of maintaining the validity of the test scores. In AQuAA, item quality is quantified across four categories of item performance statistics: item difficulty, item discrimination, item slowness (response time), and differential item functioning (DIF). Tracking these statistics could help us develop expectations about the item bank with respect to item performance, flag items with extreme and/or inadequate performance, and detect drift in measures of performance across time. Each category of item performance statistics can be computed with various statistical methods, either descriptive or model-based. For example, the item difficulty can be represented with the percentage of test takers who respond to the item correctly (i.e., descriptive method) or with the difficulty parameter estimate from an Item Response Theory (IRT) model (i.e., model-based method).

2.5 Item Exposure

The item exposure statistics concern how frequent each item (or each group of items) are used. An item being used either too frequently (over-exposure) or too infrequently (under-exposure) are undesirable for maintaining the item quality. An important statistic in this category is the item exposure rate, which is calculated as the number of test administrations containing a certain item divided by the total number of test administrations. Tracking the item exposure rates can help flag under- or over-exposure of items. Since item exposure rates are jointly affected by various factors, including the item bank characteristics and the item selection algorithm, monitoring the item exposure rate could also reveal potential issues in the item bank and the item selection algorithm.

3 How Can the Statistics Be Continuously Monitored and Updated?

To continuously monitor and update the statistics and the control charts in AQuAA, we developed an automated updating pipeline as shown in Fig. 1, and implemented it with R (R Development Core Team, 2013). The pipeline is scheduled to import data into R from a database that stores all of the assessment data (e.g., person-level data, item-response-level data and process data) daily. It then calculates the statistics that have been planned to track as described in the previous section. Before calculating the statistics, a series of automatic data inspections are conducted to check the completeness and quality of the data. For example, one item on the checklist is whether a specific day's data are missing from the datasets and whether there are any irregular values (e.g., negative values in time duration variables) in the datasets. If one or more of these tests were to fail, the pipeline would raise an alert and further investigation would be conducted by analysts to examine the causes of

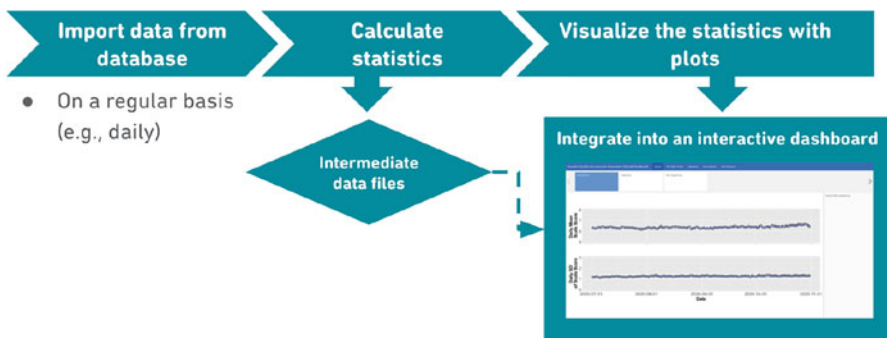


Fig. 1 QA system automated updating pipeline

the data quality issues. These data quality inspections serve to ensure the accuracy of the calculated statistics.

To visualize the trends and patterns of the statistics and facilitate communication in the QA process, statistics are plotted using the `ggplot2` R package (Wickham, 2011). Line plots are one of the most basic tools to visualize the time-series data. Smoothed lines created by the locally weighted scatterplot smoothing (LOWESS; Cleveland, 1979) method are used to represent the trends of the statistics. The statistics and figures are integrated into an interactive dashboard using the Flexdashboard (Iannone et al., 2020) package.

Some intermediate data files also result from the statistical calculation; links to these data files are also integrated into the interactive dashboard so that they can be accessed by internal researchers for quality audit and/or further analyses.

4 What Criteria Should Be Used to Identify Irregularities in the Statistics?

Since it is unfeasible to manually monitor the QA dashboard around the clock, it is necessary to have an alert system to automatically push notifications to the test developing team whenever there is an irregularity in the tracked statistics. Therefore, the third design question concerns the identification of irregularities in the statistics, which involves developing AQuAA's warning mechanism. Developing a warning mechanism in a digital-first assessment is challenging partly due to the fact that the population of test takers is constantly evolving and changing. Accordingly, the baselines of many of the tracked metrics cannot be assumed to be stationary over time. To the changing score baselines and flag daily irregularities (i.e., days when the score statistics have irregular trends), a daily score residual (r) is computed:

$$r = X_{obs} - X_{pred}$$

where X_{pred} is the daily predicted score that is computed by regressing historical scores on various background variables, such as gender, native language, and test taking intent, using a nonparametric regression tree method. The predicted score has taken into account the changing test-taker demographics and will serve as the baseline for flagging daily irregularities. X_{obs} is the daily average observed score. If a day's daily score residual has a large absolute value compared to the corresponding 30-day rolling average, (i.e., $z > 2.5$, where $z = (r - M_{30})/s_{30}$; M_{30} and s_{30} are the 30-day rolling mean and standard deviation of the daily residuals, respectively), then the daily residual is flagged as irregular. In such a context, "irregularity" is defined as a significant daily score change that cannot be explained by changes in the test taker demographics. Such irregularities should trigger alerts because they could indicate the occurrence of some unusual events. In Fig. 2, an irregular daily residual is represented as a colored point (green, blue or red, depending on severity).

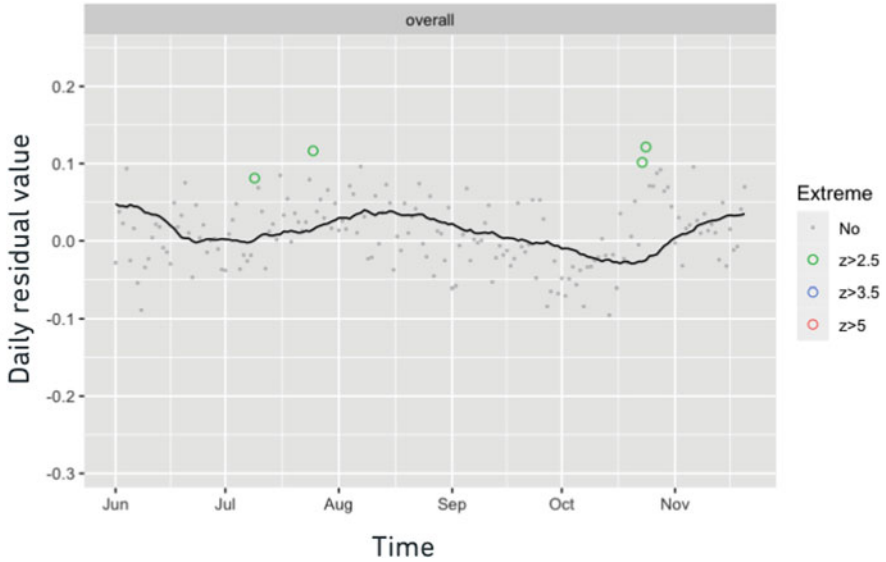


Fig. 2 Daily residual chart. Each dot represents an average daily residual. The black line represents the 30-day rolling average of the daily residuals, M_{30} . Irregular daily residuals are represented as colored dots. Green, blue and red indicate slight, medium and large severity, respectively

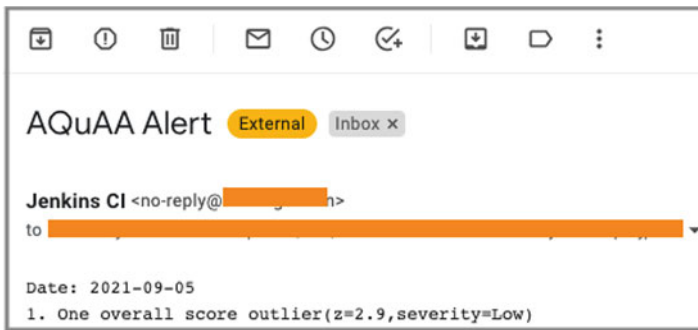


Fig. 3 Demo of an alarming email sent to the test developing team

The green points in Fig. 2 indicate that there are a few significant increases in the overall scores in July and October, but these score increases cannot be fully explained by the population changes that are indicated by the observed background variables. An automated email sent to the test developing team is triggered (See Fig. 3 for a demo) whenever a daily irregularity is detected (e.g., when there is a colored point coming up in Fig. 2). In a situation such as the one described in Fig. 2, SMEs will meet and discuss the potential causes for these outliers. One hypothesis is seasonality, but for a young test like the Duolingo English Test it is

still too early to establish it. One way to determine the causes is to explore the potential seasonality is to review major events that concur with the green points and whether we encountered a similar pattern in the previous years (although this does not work well for a new test). Examples of major events include institutions' admission deadlines, which could impose test-taking population changes that have not been fully captured by the observed background variables. For instance, further investigations reveal that the green points in late October in Fig. 2 approximately concur with many universities' early action deadline on Nov 1 (i.e., deadline for students to submit their applications early). Another hypothesis is that abnormal test taking behaviors (e.g., item preknowledge) could be another cause for the score alerts, which could further threaten the score validity. Therefore, it is important to collaborate with test security experts in understanding the causes of the alerts. In this particular case presented in Fig. 2 the evidence supports that seasonality may exist and we had no evidence of cheating or inappropriate behaviors.

5 What Actions Should Be Taken When There Are Alerts of Irregularities?

The last design question involves how actions are informed by the alerts of irregularities. Even though most of the processes (e.g., test development, test scoring) in the digital-first assessment have been automated, the QA process is still a combination of automatic processes and human review. AQuAA serves as a starting point for the human review process, and the human review process, in turn, helps AQuAA to evolve into a more powerful tool to detect assessment validity issues. Figure 4 demonstrates an example of the human review process following each week's updates of the QA statistics: SMEs meet to review the alerts raised by AQuAA's alarming mechanism and check for any anomalies that have been suggested by the quality control statistics/charts but have not been caught by AQuAA alarming mechanism. The SMEs review each individual alert and investigate the possible causes of the alert. As described in the example of Fig. 2 in Sect. 4, reviewing the major events that concur with the alerts is an important step in discovering the causes. Additionally, it is crucial to distinguish the alerts caused by seasonality from those caused by other factors that may threaten the score validity (e.g., abnormal test taking behaviors). Collaborating with experts with diverse backgrounds and expertise (e.g., test security experts, institution engagement experts) can help develop a more comprehensive picture of the events concurring with the alerts and provide more accurate insights about the causes of the alerts.

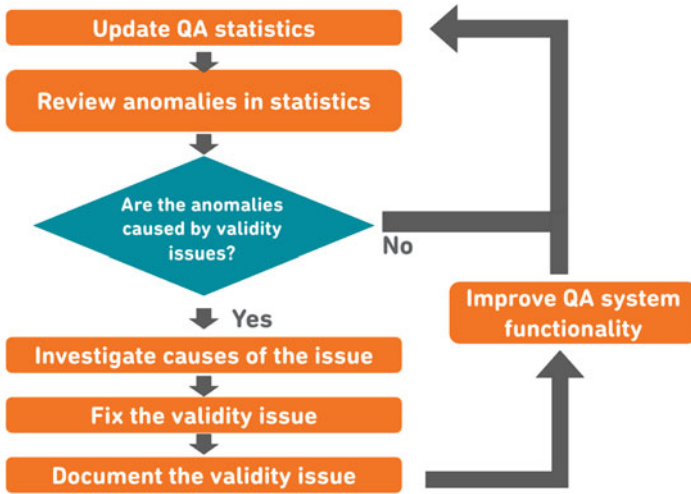


Fig. 4 Human review process in the QA procedure

If the alert is believed to be caused by a validity issue, follow-up actions are taken to determine the urgency of the issue, to fix it, and to document it. If a validity issue had not been caught by AQuAA alarming mechanism, improvements would be made to AQuAA functionality so that it would be more sensitive in detecting the issue in the future.

6 Discussion

This paper describes the development of a QA system designed for a high-stakes, digital-first language assessment. The design of AQuAA was largely motivated by the characteristics of the assessment. The key design questions and considerations have been elaborated. It should be noted that the list of QA statistics presented in this paper is not exhaustive. Instead, due to the data-rich nature of the digital-first assessment, the list of monitoring statistics is expected to be lengthened and improved as the research in statistical techniques advances. If applied to other assessments, the list of monitoring statistics should also be customized to the purposes and characteristics of the assessments. Hence, the infrastructure of AQuAA is designed to be flexible so that it can incorporate and monitor additional statistics.

In the future, besides controlling the quality of the assessment, we aim to use AQuAA to understand the seasonality of the test. From the statistics and control charts (e.g., Fig. 2), we have not seen a clear seasonality pattern yet. The lack of seasonality is possibly due to the fact that the Duolingo English Test is relatively new so that it has not accumulated sufficient data to reveal a seasonal pattern and

that the test taking population is constantly changing because of some unpredictable factors (e.g., the pandemic and the more widespread adoption of the test for institutional admission decisions). It is hoped that the seasonality pattern gets clearer as more and more test data get collected.

References

- Allalouf, A. (2007). Quality control procedures in the scoring, equating, and reporting of test scores. *Educational Measurement: Issues and Practice*, 26(1), 36–46.
- Allalouf, A., Gutentag, T., & Baumer, M. (2017). Quality control for scoring tests administered in continuous mode: An NCME instructional module. *Educational Measurement: Issues and Practice*, 36(1), 58–68. <https://doi.org/10.1111/emip.12140>
- Burstein, J., LaFlair, G., Kunnan, A., & von Davier, A. (2021). *A theoretical assessment ecosystem for a digital-first assessment—The Duolingo English Test*. <https://duolingo-papers.s3.amazonaws.com/other/det-assessment-ecosystem.pdf>
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368), 829–836.
- Haberman, S. J., Guo, H., Liu, J., & Dorans, N. J. (2008). Consistency of SAT[®] I: Reasoning test score conversions. *ETS Research Report Series*, 2008(2), i–20.
- Holland, P. W. (1994). Measurements or contests? Comments on Zwick, Bond and Allen/Donoghue. In *Proceedings of the social statistics section of the American Statistical Association* (pp. 27–29). American Statistical Association.
- Iannone, R., Allaire, J. J., Borges, B., RStudio, IO, K., Almsaeed, A., Mosbech, J., Bossart, N., Verou, L., Baranovskiy, D., Labs, S., Djuricic, B., Sardyha, T., Lewis, B., Sievert, C., Kunst, J., Hafen, R., Rudis, B., & Cheng, J. (2020). *flexdashboard: R Markdown Format for Flexible Dashboards (0.5.2)* [Computer software]. <https://CRAN.R-project.org/package=flexdashboard>
- International Test Commission. (2014). ITC guidelines on quality control in scoring, test analysis, and reporting of test scores. *International Journal of Testing*, 14(3), 195–217. <https://doi.org/10.1080/15305058.2014.918040>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
- Lee, Y.-H., & Haberman, S. J. (2013). Harmonic regression and scale stability. *Psychometrika*, 78(4), 815–829.
- Lee, Y.-H., & von Davier, A. A. (2013). Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. *Psychometrika*, 78(3), 557–575. <https://doi.org/10.1007/s11336-013-9317-5>
- Li, D., Li, S., & von Davier, A. A. (2009). Applying time-series analysis to detect scale drift. In *Statistical models for test equating, scaling, and linking* (pp. 327–346). Springer.
- Mohadjer, L., & Edwards, B. (2018). Paradata and dashboards in PIAAC. *Quality Assurance in Education*, 26(2), 263–277.
- Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2), 100–115.
- R Development Core Team. (2013). *R: A language and environment for statistical computing*. Retrieved from <http://cran.fiocruz.br/web/packages/dpIR/vignettes/timeseries-dpIR.pdf>
- Schafer, W. D., Coverdale, B. J., Luxenberg, H., & Ying, J. (2011). Quality control charts in large-scale assessment programs. *Practical Assessment, Research, and Evaluation*, 16(1), 15.
- Settles, B., LaFlair, G. T., & Hagiwara, M. (2020). Machine learning-driven language assessment. *Transactions of the Association for Computational Linguistics*, 8, 247–263. https://doi.org/10.1162/tacl_a_00310

- von Davier, A. A. (2015). *Virtual and collaborative assessments: Examples, implications, and challenges for educational measurement*. Invited Talk at the Workshop on Machine Learning for Education, International Conference of Machine Learning 2015
- von Davier, A. A. (2017). Computational psychometrics in support of collaborative educational assessments. *Journal of Educational Measurement*, 54(1), 3–11.
- von Davier, A. A., Mislevy, R., & Hao, J. (Eds.). (2021). *Computational psychometrics: New methodologies for a new generation of digital learning and assessment: With examples in R and Python*. Springer. <https://doi.org/10.1007/978-3-030-74394-9>
- Wanjohi, R. G., van Rijn, P. W., & von Davier, A. A. (2013). A state space approach to modeling IRT and population parameters from a long series of test administrations. In *New developments in quantitative psychology* (pp. 115–132). Springer.
- Wickham, H. (2011). ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(2), 180–185.

Effects of Restoring Missing Data on Uniform Differential Item Functioning Analysis



Ya-Hui Su and Bin-Chin Lu

Abstract The fairness of a test is a critical indicator of its quality and the basis on which it can confidently be administered. The detection of differential item functioning (DIF) is a vital procedure for examining test fairness. If the missing data rate is too high, DIF detection cannot function properly. Therefore, knowing how to handle missing values to ensure reliable DIF detection results is crucial. Because the method of handling missing values determines the values of the restored data, it may also subsequently affect the DIF detection results. Finch (Appl Meas Educ 24:281–301, 2011) reported that using multiple imputation (MI) to manage missing data is advantageous. However, that study only considered the target item missing and the percentage of missing responses to it instead of also incorporating the effect of the DIF percentage on the DIF detection results. Thus, this study investigated how these factors affected the restoration of missing values and the detection of DIF items. The results demonstrated that although the type of missing data had the greatest effect on data restoration, the missing data rate also affected it. The *k*-nearest neighbors (*k*NN) method restored missing completely at random (MCAR) and missing at random (MAR) data most accurately, and the predictive mean matching (PMM), MI, and classification and regression tree (CART) methods restored missing not at random (MNAR) data most accurately. The Lord's chi-square and Mantel–Haenszel DIF detection methods both met the model expectation of the type I error rate when the DIF percentage was less than 20%. As the DIF percentage increased, both methods' type I error rates increased, and their statistical power decreased.

Keywords Missing data · DIF · Multiple imputation · Mantel-Haenszel statistic · Lord's chi-square

Y.-H. Su (✉) · B.-C. Lu

Department of Psychology, National Chung Cheng University, Minhsiung, Chiayi, Taiwan
e-mail: psyys@ccu.edu.tw

1 Introduction

The fairness of a test is a critical indicator of its quality and the basis on which it can confidently be administered. The detection of differential item functioning (DIF) is a vital procedure for examining test fairness. This study examined the impact of missing item response data on the detection of uniform DIF. Uniform DIF exists when members of one group have a consistent advantage in correctly responding to an item compared with members of another group across all levels of the skill that the instrument is measuring (Camilli & Shepard, 1994). Many DIF detection methods are available, including those based on item response theory (IRT) and non-IRT approaches. Non-IRT approaches, such as the Mantel–Haenszel method (MH; Mantel & Haenszel, 1959), logistic regression (Swaminathan & Rogers, 1990), and SIBTEST (Millsap & Everson, 1993; Shealy & Stout, 1993), do not make any assumption on the distribution of the latent trait. An IRT approach, however, assumes that the latent trait is normally distributed among the population, and the item responses are assumed to follow the distribution of some IRT models. Item parameters are estimated based on the models and are then compared to identify any significant difference between groups. IRT approaches include Lord's chi-square test (Lord, 1980) and the likelihood ratio test (Thissen et al., 1993).

Rubin (1976) defined three types of missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR data have no systematic mechanism associated with the missingness, and MAR means that the probability of a missing response is the same only within groups defined by the observed data; therefore, examinees from one group might have a greater likelihood of leaving an item unanswered than examinees from another group. MNAR means that the reason for the missing data is related to factors unknown to the researcher. Peugh and Enders (2004) stated that missing data analyses have received considerable attention in the methodological literature of statistical modeling, and multiple imputation (MI; Rubin, 1987) and maximum likelihood estimation (MLE; Dempster et al., 1977) are recommended. The predictive mean matching (PMM; Little & Rubin, 2002) method is also a common method for restoring missing data. Missing data can be deemed a classification problem in data mining, which is the process of extracting and discovering patterns in large data sets. The classification and regression tree (CART; Breiman et al., 1984) and k -nearest neighbors (k NN; Fix & Hodges, 1951) algorithms are commonly used in data mining. However, the CART and k NN methods have not been applied to restore missing data. Methods based on statistic modeling and data mining are included in this study to restore missing data.

Some DIF detection methods have been applied to missing item response data (Finch, 2008, 2011; Huisman & Molenaar, 2001; Sijtsma & van der Ark, 2003). If the missing data rate is too high, however, DIF detection cannot function properly. Therefore, knowing how to manage missing values to ensure reliable DIF detection results is essential. Because the method of handling missing values determines the values of the restored data, it may subsequently affect the DIF detection results.

Finch (2011) reported that using MI to manage missing data is advantageous. However, Finch only considered the target item missing and the percentage of missing responses for it rather than incorporating the effect of the DIF percentage on the DIF detection results. Robitzsch and Rupp (2009) indicated that interactions among the type of missing data, data restoration methods, and missing rates might greatly affect predictions of DIF detection efficiency. Thus, this study investigated how these factors affect the restoration of missing values and the detection of DIF items.

2 Method

Two simulations were conducted, and data were generated on the basis of the Rasch (1960) model. The sample size was 500 for both the reference and focal groups. The latent traits of both groups were drawn from the standard normal distribution, and 40 item parameters from Robitzsch and Rupp (2009) were used.

2.1 Study 1

The first study focused on data restoration, and five variables were manipulated: the type of missing data (three levels: MCAR, MAR, and MNAR), missing data rate (four levels: 0%, 10%, 20%, and 30%), data restoration method (five types: MI, PMM, MLE, CART, and k NN), DIF percentage (five levels: 0%, 10%, 20%, 30%, and 40%), and DIF amount (three levels: 0, 0.5 and 0.8). Each condition was replicated 100 times.

For each of the types of missing data conditions, the missing data rate determined the number of responses that were randomly selected to be missing for each group. Under the MCAR condition, responses from both groups were selected to be missing, and under the MAR condition, responses from only the focal group were selected to be missing. Under the MNAR condition, incorrect responses from both groups were selected to be missing.

MI allows for uncertainty regarding the missing data by creating several imputed data sets and combining the results obtained from each. PMM is an appealing method for conducting multiple imputations for missing data and reduces bias through imputation by employing real values sampled from the data. For each missing entry, the method involves forming a small set of cases with complete and similar responses, and the observed response from one randomly selected case replaces the missing entry. MLE uses available responses from each case to compute maximum likelihood estimates for the missing responses. CART seeks predictors that are used to split the sample by repeatedly dividing the sample into more homogeneous subsamples; the algorithm uses surrogate splits to calculate the best split (left or right) for a case with a missing entry if more cases are sent in the

same direction. *k*NN replaces the missing entry with the mean value from the *k*NN parameter based on training. In this study, a Gower distance (Gower, 1971) was used to impute missing values. Evaluation criteria included five indicators, namely accuracy, bias, root mean squared error, mean absolute error, and correlation.

2.2 Study 2

The second study focused on DIF detection, and six variables were manipulated. In addition to the variables used in Study 1, DIF detection methods (two levels: MH and Lord's chi-square) were also incorporated. Each condition was replicated 100 times.

The MH method uses test scores as matching variables for the reference and focal groups and calculates the common odds ratio for each item to determine whether an item has DIF. The Lord's chi-square method entails computing item parameters for the reference and focal groups and then determining whether the differences in item parameters are statistically significant. Although the MH method is a non-IRT approach and Lord's chi-square method is an IRT approach, both methods are commonly used for DIF detection. Evaluation criteria included two indicators: type I error and statistical power. The results were calculated over 100 replications.

3 Results

3.1 Study 1

The results from Study 1 of the MCAR, MAR, and MNAR data are listed in Tables 1, 2, and 3, respectively. When the missing data rates were 10%, 20%, or 30%, the data restoration accuracy should have started at 0.9, 0.8, and 0.7, respectively. The type of missing data had a great effect on the data restoration accuracy; the accuracy of restoring MCAR data was slightly higher than for MAR data, which was higher than for MNAR data. When MCAR and MAR data had a 10% missing data rate, the *k*NN method performed best of all the methods, and the MLE method performed worst. When MNAR data had a 10% missing rate, the PMM, MI, and CART methods performed best, and the *k*NN method performed worst. With all data, the PMM, MI, and CART methods performed similarly, and the DIF percentage and DIF amount had little effect on the accuracy of the data restoration. Data restoration was less accurate as missing data rates increased.

Table 1 Data restoration accuracy rates for MCAR data

Missing rate (%)	DIF amount	DIF%	PMM	MI	CART	kNN	MLE	
10	0	0	0.964	0.964	0.964	0.968	0.955	
	0.5	10	0.964	0.964	0.963	0.967	0.955	
		20	0.963	0.963	0.963	0.966	0.955	
		30	0.963	0.963	0.963	0.967	0.955	
		40	0.962	0.962	0.962	0.966	0.956	
	0.8	10	0.963	0.963	0.963	0.967	0.955	
		20	0.964	0.964	0.963	0.967	0.955	
		30	0.963	0.963	0.963	0.966	0.956	
		40	0.964	0.964	0.964	0.968	0.955	
	20	0	0	0.928	0.929	0.928	0.935	0.911
		0.5	10	0.927	0.927	0.927	0.933	0.911
			20	0.926	0.926	0.925	0.931	0.911
30			0.927	0.927	0.927	0.933	0.911	
40			0.925	0.925	0.924	0.931	0.912	
0.8		10	0.927	0.927	0.927	0.933	0.911	
		20	0.927	0.927	0.927	0.933	0.911	
		30	0.926	0.926	0.926	0.931	0.913	
		40	0.928	0.928	0.928	0.934	0.911	
30		0	0	0.893	0.893	0.892	0.901	0.866
		0.5	10	0.891	0.891	0.890	0.898	0.867
			20	0.888	0.888	0.888	0.894	0.868
	30		0.891	0.890	0.890	0.899	0.867	
	40		0.887	0.887	0.887	0.894	0.868	
	0.8	10	0.890	0.890	0.890	0.898	0.867	
		20	0.891	0.891	0.891	0.898	0.866	
		30	0.889	0.889	0.889	0.894	0.870	
		40	0.892	0.892	0.891	0.900	0.866	

3.2 Study 2

The second study focused on DIF detection. Because MAR data performed poorly and MCAR performed similarly to MNAR, only the MCAR results are discussed. The type I error and statistical power results of the two DIF detection methods for a 0% missing data rate are listed in Table 4. The results of the type I error and statistical power of the MH and Lord’s chi-square methods using MCAR data with a 10% missing data rate are listed in Table 5, and the results of each method for MCAR data with a 30% missing rate are listed in Table 6. Because each condition was replicated 100 times, the average type I error and power were calculated over 100 replications. The type I error should be approximately 5% to meet the model expectations under each condition. If the type I error was less than 4% or more than 6%, the result and the corresponding power in Tables 4, 5, and 6 appear in bold.

Table 2 Data restoration accuracy rates for MAR data

Missing rate (%)	DIF amount	DIF%	PMM	MI	CART	kNN	MLE	
10	0	0	0.964	0.964	0.964	0.966	0.955	
	0.5	10	0.964	0.964	0.963	0.965	0.955	
		20	0.963	0.963	0.963	0.964	0.955	
		30	0.963	0.963	0.963	0.965	0.955	
		40	0.963	0.962	0.962	0.964	0.955	
	0.8	10	0.964	0.964	0.963	0.965	0.955	
		20	0.964	0.964	0.964	0.965	0.955	
		30	0.963	0.963	0.963	0.964	0.956	
		40	0.964	0.964	0.964	0.965	0.955	
	20	0	0	0.928	0.928	0.928	0.931	0.910
		0.5	10	0.927	0.927	0.926	0.929	0.910
			20	0.925	0.925	0.925	0.925	0.911
30			0.926	0.926	0.926	0.928	0.910	
40			0.924	0.924	0.924	0.925	0.911	
0.8		10	0.927	0.927	0.926	0.929	0.910	
		20	0.927	0.927	0.927	0.928	0.910	
		30	0.926	0.926	0.926	0.925	0.912	
		40	0.928	0.928	0.927	0.929	0.910	
30		0	0	0.890	0.890	0.890	0.896	0.865
		0.5	10	0.889	0.889	0.888	0.893	0.865
			20	0.886	0.886	0.886	0.887	0.865
	30		0.888	0.888	0.887	0.892	0.866	
	40		0.885	0.885	0.884	0.887	0.866	
	0.8	10	0.888	0.889	0.888	0.894	0.866	
		20	0.889	0.890	0.889	0.893	0.864	
		30	0.888	0.888	0.887	0.887	0.867	
		40	0.890	0.890	0.889	0.895	0.864	

The overall results of Study 2 were highly similar for the 0%, 10% and 30% missing data rates. For all the rates displayed, the type I error of Lord’s chi-square method was slightly lower than the model expectation when the DIF percentage was either 0% or 10% with a 0.5 DIF amount. The type I error of Lord’s chi-square method met the model expectation when the DIF percentage was 10% with a 0.8 DIF amount or when the DIF percentage was 20% with a 0.5 DIF amount. As the DIF amount increased, the type I error and the power both increased. As the DIF percentage increased, the type I error increased but power decreased. Multiple data restoration methods performed similarly in terms of type I error and power.

Table 3 Data restoration accuracy rates for MNAR data

Missing rate (%)	DIF amount	DIF%	PMM	MI	CART	kNN	MLE	
10	0	0	0.960	0.960	0.960	0.936	0.951	
	0.5	10	0.960	0.960	0.959	0.935	0.951	
		20	0.958	0.958	0.958	0.934	0.951	
		30	0.960	0.960	0.959	0.935	0.951	
		40	0.958	0.958	0.958	0.933	0.951	
	0.8	10	0.960	0.960	0.959	0.935	0.951	
		20	0.960	0.960	0.959	0.936	0.950	
		30	0.958	0.958	0.958	0.936	0.951	
		40	0.960	0.960	0.960	0.936	0.951	
	20	0	0	0.911	0.911	0.910	0.849	0.890
		0.5	10	0.910	0.910	0.909	0.846	0.891
			20	0.906	0.906	0.905	0.837	0.889
30			0.910	0.910	0.909	0.846	0.891	
40			0.906	0.906	0.905	0.839	0.891	
0.8		10	0.909	0.909	0.909	0.845	0.891	
		20	0.909	0.909	0.908	0.843	0.889	
		30	0.905	0.904	0.904	0.834	0.888	
		40	0.910	0.910	0.910	0.847	0.890	
30		0	0	0.845	0.845	0.843	0.751	0.812
		0.5	10	0.843	0.843	0.841	0.747	0.813
			20	0.834	0.834	0.831	0.728	0.807
	30		0.843	0.843	0.841	0.749	0.814	
	40		0.836	0.836	0.833	0.737	0.812	
	0.8	10	0.842	0.842	0.840	0.747	0.813	
		20	0.839	0.839	0.837	0.738	0.807	
		30	0.830	0.830	0.828	0.718	0.803	
		40	0.844	0.844	0.841	0.748	0.812	

When the missing data rate was 0%, the type I error of the MH method was slightly lower than the model expectation when the DIF percentage was 0%, and when the missing data rate was 10%, the type I error of the MH method met the model expectation when the DIF percentage was 0% or when the DIF percentage was 10% with a 0.5 DIF amount. When the missing rate reached 30%, however, the type I error of the MH method was slightly lower than the model expectation when the DIF percentage was 0%. The type I error of the MH method met the model expectation when the DIF percentage was 10% with a 0.5 DIF amount. The type I error of Lord's chi-square method was slightly smaller than the model expectation when the DIF percentage was either 0% or 10% with a 0.5 DIF amount. As the missing data rate increased, the type I error and power decreased.

Table 4 Type I error and power of 0% missing rate

DIF methods	DIF amount	DIF%	Type I error	Power	
MH	0	0	.03		
	0.5	10	.04	.84	
		20	.08	.74	
		30	.13	.57	
		40	.23	.49	
	0.8	10	.06	.99	
		20	.16	.98	
		30	.31	.93	
		40	.50	.84	
	Lord	0	0	.02	
		0.5	10	.03	.82
			20	.06	.72
30			.11	.54	
40			.21	.44	
0.8		10	.05	.99	
		20	.12	.99	
		30	.29	.92	
		40	.48	.82	

Table 5 Type I error and power with 10% missing rate of MCAR data

DIF methods	DIF amount	DIF%	Type I error					Power					
			PMM	MI	CART	KNN	MLE	PMM	MI	CART	kNN	MLE	
MH	0	0	.04	.03	.04	.04	.03						
	0.5	10	.04	.04	.04	.04	.04	.73	.73	.74	.73	.72	
		20	.08	.08	.08	.07	.08	.65	.65	.64	.65	.63	
		30	.12	.12	.12	.11	.11	.49	.49	.49	.50	.46	
		40	.20	.19	.20	.20	.19	.40	.41	.42	.41	.38	
	0.8	10	.06	.06	.06	.06	.06	.98	.98	.98	.99	.98	
		20	.15	.15	.14	.15	.14	.95	.95	.95	.96	.95	
		30	.27	.27	.26	.27	.25	.88	.88	.89	.90	.86	
		40	.43	.43	.43	.42	.41	.76	.76	.76	.77	.72	
	Lord	0	0	.03	.02	.03	.02	.02					
		0.5	10	.03	.03	.03	.03	.03	.71	.71	.70	.72	.72
			20	.05	.05	.05	.05	.06	.64	.63	.63	.64	.62
30			.09	.10	.10	.10	.08	.45	.46	.45	.46	.43	
40			.18	.17	.17	.16	.16	.37	.37	.37	.38	.35	
0.8		10	.04	.04	.04	.04	.04	.98	.98	.98	.99	.98	
		20	.11	.11	.11	.11	.10	.95	.96	.95	.96	.94	
		30	.25	.25	.24	.25	.22	.86	.87	.88	.88	.84	
		40	.41	.41	.41	.42	.38	.74	.75	.75	.76	.70	

Table 6 Type I error and power with 30% missing rate of MCAR data

DIF methods	DIF amount	DIF%	Type I error					Power					
			PMM	MI	CART	KNN	MLE	PMM	MI	CART	kNN	MLE	
MH	0	0	.03	.03	.03	.03	.03						
	0.5	10	.04	.04	.04	.04	.04	.48	.51	.51	.55	.48	
		20	.07	.06	.06	.06	.05	.44	.43	.43	.44	.40	
		30	.08	.08	.08	.08	.07	.32	.33	.32	.35	.28	
		40	.13	.13	.13	.14	.11	.25	.26	.26	.27	.23	
	0.8	10	.06	.06	.05	.06	.05	.88	.90	.88	.91	.89	
		20	.12	.11	.10	.11	.08	.80	.81	.80	.81	.78	
		30	.20	.19	.17	.18	.14	.71	.71	.71	.72	.64	
		40	.31	.30	.27	.30	.24	.58	.57	.53	.57	.46	
	Lord	0	0	.02	.02	.02	.03	.02					
		0.5	10	.03	.03	.03	.03	.03	.46	.48	.48	.50	.44
			20	.06	.05	.05	.04	.04	.42	.41	.39	.42	.36
30			.07	.07	.07	.07	.06	.28	.29	.28	.31	.25	
40			.11	.11	.11	.12	.10	.22	.23	.23	.23	.19	
0.8		10	.04	.04	.04	.04	.04	.89	.90	.88	.90	.88	
		20	.09	.09	.08	.08	.07	.80	.80	.81	.81	.74	
		30	.18	.17	.16	.16	.12	.67	.68	.68	.68	.60	
		40	.28	.28	.25	.28	.20	.56	.55	.53	.57	.44	

4 Conclusions and Discussion

Four conclusions were drawn from this study. First, the DIF percentage and DIF amount had little effect on the accuracy of data restoration but had a great effect on DIF detection. Second, the type and rate of the missing data had a great effect on the accuracy of data restoration and DIF detection. Third, the kNN method restored MCAR and MAR data most accurately, and the PMM, MI, and CART methods restored MNAR data most accurately. However, among these three methods, the MI method had a longer computation time than the PMM or CART methods. Fourth, Lord’s chi-square and the MH DIF detection methods both met the model type I error expectations when the DIF percentage was less than 20%. As the DIF percentage increased, the type I error rates of both methods increased, and the statistical power decreased.





Future research can focus on several topics. Similar to Finch (2011), the missing data in this study was also manipulated. However, the missing data could be manipulated in other manners depending on the type of data. This study was conducted for uniform DIF conditions, according to the Rasch model. In practice, nonuniform DIF conditions are common, and therefore, other models could be considered for future studies. Five data restoration methods were manipulated in this study, but other methods could be considered in the future. The MH and Lord’s chi-square DIF detection methods were used in the study, but other methods of purification might improve the efficacy of DIF detection. For example, a two-stage or iterative MH method should be more effective than a one-stage MH method.

References

- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth and Brooks.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage Publications.
- Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1–38.
- Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Education Measurement*, 45(3), 225–245.
- Finch, H. (2011). The use of multiple imputation for missing data in uniform DIF analysis: Power and type I error rates. *Applied Measurement in Education*, 24(4), 281–301. <https://doi.org/10.1080/08957347.2011.607054>
- Fix, E., & Hodges, J. L. (1951). *Discriminatory analysis: nonparametric discrimination, consistency properties*. USAF School of Aviation Medicine.
- Gower, J. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27(4), 857–871. <https://doi.org/10.2307/2528823>
- Huisman, M., & Molenaar, I. W. (2001). Imputation of missing scale data with item response models. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 221–244). Springer.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. John Wiley & Sons.
- Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Lawrence Erlbaum Associates.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4), 719–748. <https://doi.org/10.1093/jnci/22.4.719>
- Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, 17(4), 297–334.
- Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74(4), 525–556. <https://doi.org/10.3102/00346543074004525>
- Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.
- Robitzsch, A., & Rupp, A. A. (2009). Impact of missing data on the detection of differential item functioning: The case of Mantel-Haenszel and logistic regression analysis. *Educational and Psychological Measurement*, 69(1), 18–34. <https://doi.org/10.1177/0013164408318756>
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. <https://doi.org/10.1093/biomet/63.3.581>
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.
- 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(2), 159–194. <https://doi.org/10.1007/BF02294572>
- Sijtsma, K., & van der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research*, 38, 505–528.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361–370. <https://doi.org/10.1111/j.1745-3984.1990.tb00754.x>
- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Lawrence Erlbaum Associates.

Two-Step Approach to Topic Modeling to Incorporate Covariate and Outcome



Minju Hong , Hye-Jeong Choi, Constanza Mardones-Segovia ,
Yasemin Copur-Gencturk , and Allan S. Cohen 

Abstract This study investigates the applicability of topic modeling to analyze educational data. Topic modeling is useful because it reveals the latent topic structure underlying a collection of texts. Because metadata provides useful information about the topics, this study explores a way of including metadata as a covariate predicting topics and outcomes by topic in a topic model using a two-step approach. In Study 1, we use structural topic model (STM) and regression model because STM estimates the topic structure and how covariate is related to the topics. In Study 2, supervised Dirichlet allocation (sLDA) model is used to investigate the relationship between topics and the outcome variable: we incorporate sLDA with ANOVA. We demonstrate that the inclusion of multiple metadata improved the interpretability of the topic modeling techniques' results by examining the relationship among the examinees' written answers, problem-solving strategies, and scores using the empirical data of 246 middle school mathematics teachers' written responses to an item.

Keywords Topic modeling · Metadata · Mathematics education · Structural topic model · Supervised Dirichlet allocation

M. Hong (✉)

Department of Educational Psychology, University of Georgia, Athens, GA, USA
e-mail: mh19985@uga.edu

H.-J. Choi

The University of Georgia, Athens, GA, USA

The Human Resources Research Organization, Louisville, KY, USA

C. Mardones-Segovia · A. S. Cohen

University of Georgia, Athens, GA, USA

Y. Copur-Gencturk

University of Southern California, Los Angeles, CA, USA

1 Introduction

Topic modeling is a technique used to extract a hidden thematic structure underlying a collection of documents (Blei, 2012; Nagwani, 2015). Because topic modeling allows researchers to investigate how a person generates a document using statistical estimation, it has been studied in such diverse disciplines as computer science (Phan et al., 2008) and journalism (Paul & Dredze, 2011). In education, topic modeling has been used to investigate examinees' thinking and reasoning processes based on the text of their answers to constructed response (CR) items (Cardozo-Gaibisso et al., 2020; Kim et al., 2017; Xiong et al., 2019).

In addition to text, the documents contain additional information, such as the author's demographics, publication years, or ratings of the documents, called metadata. Because researchers have shown that the inclusion of such metadata in the model improves the interpretability of the analysis results, the models under topic modeling framework, such as supervised latent Dirichlet allocation model (sLDA) (McAuliffe and Blei, 2007) or structural topic model (STM) (Roberts et al., 2014), have been proposed.

We conducted two studies to demonstrate how to include multiple metadata for analyzing textual datasets in education. By analyzing the mathematics educational dataset, we demonstrate how teachers' written answers to a proportional reasoning item, their problem-solving strategies, and their scores were related to one another.

2 Theoretical Backgrounds

2.1 Structural Topic Model (STM)

The STM model (Roberts et al., 2014) is a topic model that includes covariates in the model in order to help guide detection of the latent thematic structure (i.e., the latent topics) in a collection of documents. Suppose that we have a collection of D documents (i.e., a corpus) with a vocabulary of size V that has K latent topics. Each document is denoted as $d \in \{1, \dots, D\}$, and the words used in each document are denoted as $n \in \{1, \dots, N_d\}$. Then, under STM, the generative process of documents is as follows:

1. For document d with topic prevalence covariate X_p , choose per-document topic proportions $\theta_d \sim \text{LogNormal}(\gamma X_p, \Sigma)$ where γ is the coefficient of topic prevalence covariate.
2. For the n th word with topical content covariate X_c , choose per-topic word probabilities $\varphi_k \sim \exp(m + \kappa_k + \kappa_{X_c} + \kappa_{k, X_c})$ where m is the marginal log-frequency of word, and κ is the coefficient of topic content covariate.

3. For the n th word in the d th document,

- (i) choose the per-word topic assignment $z_{d,n} \sim \text{Multinomial}(\theta_d)$
- (ii) for the k th topic, choose the word $w_{d,n} \sim \text{Multinomial}(\varphi_{z_{d,n}})$

2.2 Supervised Latent Dirichlet Allocation (sLDA) Model

The sLDA topic model (McAuliffe and Blei, 2007) is different from the STM as it includes metadata as an outcome variable predicted by topics rather than as covariates. For this reason, it is useful to investigate the relationship between topics and outcome. With the same notations used above, the document-generating process document under sLDA is as follows:

1. For document d , choose per-document topic proportions $\theta_d \sim \text{Dirichlet}(\alpha)$
2. For the n th word, choose per-topic word probabilities $\varphi_k \sim \text{Dirichlet}(\beta)$
3. For the n th word in the d th document,
 - (i) choose per-word topic assignment $z_{d,n} \sim \text{Multinomial}(\theta_d)$
 - (ii) for the k th topic, choose the word $w_{d,n} \sim \text{Multinomial}(\varphi_{z_{d,n}})$
4. Choose $y \sim \text{GLM}(\eta^\top \bar{z}, \sigma^2)$ where $\bar{z} := \sum_{d=1}^D \sum_{n=1}^{N_d} z_{d,n}$ where N_d is the number of words in the d th document.

3 Methods

3.1 Data Description

The data in this study consist of written responses of 246 middle-school mathematics teachers to items to assess their proportional reasoning. We examined teachers' proportional reasoning by asking them to explain the relationship between the width and the height of a photograph.

Variables for this study included the rubric-based score of the item (i.e., score), the problem-solving strategy used by the teacher (i.e., strategy), and the teacher's written answer (i.e., text). Two mathematics education experts assigned five strategies and four categories of scores to each teacher's answer. In this study, the strategy was used as a topic covariate to predict the use of topics, and the score was treated as an outcome variable of topic use.

Table 1 presents the frequency of teachers' problem-solving strategies and scores. Teachers used scale strategy ($N = 113, 45.94\%$) the most. For scores, most of the teachers obtained a score of 3, ($N = 194, 78.86\%$), and about 12% of the teachers obtained zero scores ($N = 30, 12.30\%$).

Table 1 Descriptive statistics of 246 teachers by strategy and score

Strategy	Score (Mean = 2.51, S.D. = 1.03)				
	0	1	2	3	Total (%)
Additive	26	0	0	0	26 (10.57)
Proportion	0	3	6	57	66 (26.83)
Scale	1	1	2	109	113 (45.94)
Unit	0	0	2	11	13 (5.29)
Other	3	5	3	17	28 (11.38)
Total (%)	30 (12.20)	9 (3.66)	13 (5.29)	194 (78.86)	246

Note: Number in parenthesis is the percentage of teachers; S.D. = standard deviation

3.2 Data Analysis

Before analyzing the data, we conducted a series of preprocessing steps. Because the data of this study contained many mathematics expressions in teachers' answers it was necessary to first convert them. We categorized mathematics expressions into two types. The first type was mathematical expressions with the equal sign “=”. We converted this type into their mathematical operators only. For example, the equation “ $16 + 16 + 8 = 40$ ” was changed to “eq_addition.” The reason we added “eq_” in front of the operation was to distinguish it by the unique words “addition”, “subtraction”, “multiplication”, and “division” originally used in the teacher's written answers.

The second type was mathematical expressions without an equal sign. This group had seven subtypes; we converted them by following the rules (shown below). These converted words were treated as one token: [Rule 1] the ratios and the fractions were converted to the phrases “math_ratio” and “math_fraction,” respectively; [Rule 2] the numbers with or without the variables or the unit scales were changed to be textual (e.g., $10x \rightarrow \text{ten}_x$, $10\text{cm} \rightarrow \text{ten_centimeters}$, $10 \rightarrow \text{ten}$, $0.1 \rightarrow \text{point_one}$).

After converting the mathematical expressions, removed punctuation, corrected typographical errors, and replace upper-case with lower-case letters. Then we performed tokenization, a step that breaks each sentence into individual words, and normalization, a step that converts the words into their stems according to the linguistic root., For example, we changed a plural noun to a singular form (e.g., lengths \rightarrow length) or a past tense verb to present tense (e.g., realized \rightarrow realize). We then removed the stop words. Stop words are high frequency but low information words that appear in nearly every document, but their meanings do not affect the topic structure (Schofield et al., 2017).

As shown in Table 2, the final dataset of 246 written answers contained 3253 tokens, 242 unique words, and an average document length of 15.40.

For Study 1, we used Approach 1, which was a combination of the STM and the linear regression model. In the first step, we extracted topics using the STM with the hyperparameter priors γ and Σ set as 1. This study did not consider the topic content covariate, so we did not set a prior for κ .

Table 2 Descriptive statistics of data after pre-processing

# of documents	# of tokens	# of words	Average length of document (S.D.)
246	3253	242	13.22 (9.15)

To determine the best fitting model, we used semantic coherence and exclusivity (Roberts et al., 2014). Semantic coherence is an index of how likely the words under a given topic are frequently co-occur in the same document. For a list of top- R words including the words v_i and v_j , semantic coherence is calculated by $\sum_{i=2}^R \sum_{j=1}^{i-1} \left(\frac{D(v_i, v_j) + 1}{D(v_j)} \right)$ where $D(v_j)$ is the number of documents in which the word v_j appears at least once, and $D(v_i, v_j)$ is the number of documents in which the words v_i and v_j appear in common. This study used the top 20 words to calculate semantic coherence, which were the most probable 20 words of the topic with the highest word probabilities.

Exclusivity is an index of how unique the most probable words of one of the topics is in the other topics in the model. For example, under a 3-topic model, if the top 20 words under Topic 1 show low probabilities under Topic 2, then Topic 1 and Topic 2 are considered exclusive to each other. Exclusivity is calculated by $\left(\frac{\omega}{ECDF(\varphi_{k,v} / \sum_{k=1}^K \varphi_{k,v})} + \frac{1-\omega}{ECDF(\varphi_{k,v})} \right)^{-1}$ where $ECDF$ is the empirical cumulative density function, φ is the word probability of the word v over the topic k , and ω is the weight for calculation. This study set ω as 0.7. By the definition of semantic coherence and exclusivity, the higher values of two measures indicate the better model fit.

In addition, using the STM, the coefficient estimates β_{kx} , that explain the relationship between the topics and the teachers' strategies, were derived from

$$\begin{aligned} \theta_{topic1} &= \beta_{11} \times strategy_1 + \dots + \beta_{1X} \times strategy_X \\ &\dots \\ \theta_{topicK} &= \beta_{K1} \times strategy_1 + \dots + \beta_{KX} \times strategy_X \end{aligned} \quad (1)$$

where θ_{topic_k} indicates the proportion of the k th topic over each written answer with k indicating each K topic and x indicating each X strategy.

In the second step, we used the linear regression model

$$score = \beta'_1 \times \theta_{topic1} + \dots + \beta'_K \times \theta_{topicK} \quad (2)$$

to explain how the estimated uses of topics (i.e., topic k th proportions, θ_{topic_k}) were related to the teachers' scores by comparing the coefficients β'_k .

For Study 2, we used Approach 2, which was a combination of sLDA (Blei & McAuliffe, 2008) and the ANOVA model. In the first step, we extracted topics using sLDA with the priors α as $1/K$, where K indicates the number of topics, and β as 0.1. To determine the best-fitting model, we used two indexes as follow. The first index

was proposed by Cao et al. (2009), that is, a measure of the cosine similarity of word probabilities between a pair of topics. The Cao et al. (2009) index is calculated by

$avg_{k \neq k' \in K} \left(\frac{\sum_{v=1}^V \varphi_{k,v} \varphi_{k',v}}{\sqrt{\sum_{v=1}^V (\varphi_{k,v})^2} \sqrt{\sum_{v=1}^V (\varphi_{k',v})^2}} \right)$ where there are V words with K topics with the word probabilities φ . If Cao et al. (2009)'s index is lower, which indicates the cosine similarity between two topics is smaller, that is, they are exclusive to each other, then, it means the better model fit.

The second index was the Jenson-Shannon divergence (JSD; Deveaud et al., 2014). This is a measure of word probability between a pair of topics. It is calculated by $avg_{k \neq k' \in K} \left[\left(\frac{1}{2} \sum_k \sum_v \varphi_{k,v} \log \left(\frac{\varphi_{k,v}}{\varphi_{k',v}} \right) \right) + \left(\frac{1}{2} \sum_{k'} \sum_v \varphi_{k',v} \log \left(\frac{\varphi_{k',v}}{\varphi_{k,v}} \right) \right) \right]$. When JSD is large, the word probabilities of the topics are far from each other, that is, each topic is exclusive to each other. This indicates a better model fit.

From the results of the best-fitting sLDA model, we explained the relationship between the extracted topics and the teachers' scores by the coefficients β'_k (see Eq. 2). For the second step, we used the ANOVA model to show how the estimated topic proportions, θ_{topic_k} , were different among the teachers depending on their strategies by the coefficients β_{kx} (see Eq. 1).

Study 2 used two criteria, cosine similarity and JSD, to measure how close the topics are that were extracted in Study 1 and Study 2. To deal with the topic-related information, we investigated per-topic word probabilities and per-document topic proportions. We compared the cosine similarity and JSD of the pair of the per-document topic proportions between Study 1 and Study 2.

4 Results

4.1 Study 1: Results of Approach 1, STM and Regression

To find the best-fitting model, we calculated semantic coherence and exclusivity from two- to ten-topic models. As shown in the left plot of Fig. 1, because the three-topic model showed the highest semantic coherence (-62.939) and exclusivity (8.380), thus, the best fitting model was determined to be the three-topic model.

This study interpreted the extracted three topics by considering which mathematical operations and answering strategies were used. We labeled Topic 1 as Setting Up a Proportion Strategy. Top 10 representative written answers for each topic were selected and reviewed for each topic in the model. For example,

*"I have to find the **proportion** to be able to get the answer. $16/10 = x/25$. I have to **cross-multiply** so $10x = 16x25$. $10x = 400$. $x = 40$. answer is 40 cm. to check you can check the **proportions**. $16/10 = 8/5$. $40/25 = 8/5$."*

*"The poster is 40 cm high. I set up my two **proportions**. $16/10 = x/25$, **crossed multiplied** to get $16x25 = 400$ and 10 **times** x which left me with $400 = 10x$. **Divided** both sides by 10 which got me my answer of 4."*

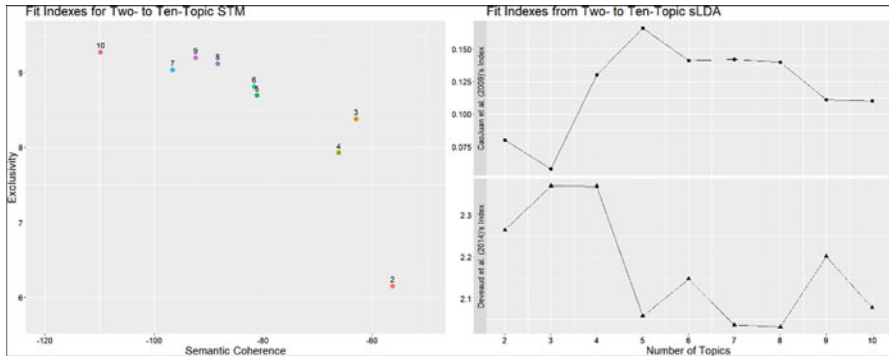


Fig. 1 Fit indexes of two- to ten-topic models for STM (left) and sLDA (right)

contained the words (in bolded) expressing multiplication operations with setting up a proportion. Also, these bolded words were among the top 20 highest probability words for Topic 1.

We interpreted Topic 2 as Use of Multiplicative Comparison Strategies. Top 10 written answers of Topic 2. For example,

*“The poster is 24-inche-high because the image is **scaled 1.5 times larger**: $1.5 \times 16 = 24$.”*

*“It is 40 cm tall because the **scale** factor from 10 cm to 25 cm is 2.5 so I **multiplied** 16 by 2.5 to get 40 cm.”*

*“The poster will be 40 cm. high. To get the width to be 25 cm. the **scale** needed to be raised by 2.5. 16 **multiplied** by 2.5 is 40.”*

contained the word “scale” and used the multiplication equations frequently.

Topic 3 was interpreted as Use of Additive Comparison Strategies. Examples of the top written answers for Topic 3 included the following:

*“To get an estimate I would take 25 and **subtract** 10 from it which would 15. So, the width is 15cm more. **Adding** 15cm to the original height would make it 31. Another way to do this is to see that on the photograph the width is 6 cm **smaller** than the height. So, on the poster **add** 6cm to the width and you would get 31 cm as the height.”*

*“The picture is 31 cm high. I **added** 15 to the high because that is how much **bigger** the width is from the first picture.”*

*“The poster is 31cm high. I got my answer because I found the difference in width of the two posters ($25 - 10 = 15$). Then I took that solution of 15 and **added** it to the width of the first poster (16cm). I know that $15\text{cm} + 16\text{cm} = 31\text{ cm}$ for the width.”*

These all used equations with addition and the words related to additive strategy.

Table 3 shows the results explaining the relation between strategies and individual topics. At the $\alpha = 0.05$ level, proportion strategy was the largest significant factor for Topic 1 ($\beta_{1_proportion} = 0.74$ with $p < .001$); thus, the interpretation of Topic 1 was reasonable. For Topic 2, because the estimated coefficient of scale strategy ($\beta_{2_scale} = 0.80$ with $p < .001$) was the largest, this strategy was the

Table 3 Effects of problem-solving strategies on topics from STM

Strategy	Estimate	S.E.	t-statistic	p-value
Topic 1 Setting up a proportion				
Additive	0.04	0.03	0.14	0.89
Proportion	0.74	0.03	29.19	<.001
Scale	0.17	0.02	8.39	<.001
Unit	0.28	0.06	4.48	<.001
Other	0.10	0.03	2.86	<.001
Topic 2 Use of multiplicative comparison strategies				
Additive	0.01	0.03	0.47	0.64
Proportion	0.25	0.03	9.87	<.001
Scale	0.80	0.02	40.08	<.001
Unit	0.70	0.06	11.28	<.001
Other	0.77	0.03	23.37	<.001
Topic 3 Use of additive comparison strategies				
Additive	0.98	0.00	313.35	<.001
Proportion	0.01	0.00	3.13	<.001
Scale	0.03	0.00	19.30	<.001
Unit	0.01	0.00	3.19	<.001
Other	0.13	0.01	21.47	<.001

S.E. = standard error

Table 4 Effects of topic uses on teachers' scores from regression

Topic	Est.	S.E.	t-stat.	p-value
Topic 1 Setting up a proportion strategy	2.75	0.08	33.77	<.001
Topic 2 Use of multiplicative comparison strategies	3.02	0.05	56.79	<.001
Topic 3 Use of additive comparison strategies	0.09	0.10	0.86	0.39

S.E. = standard error

most meaningful factor. For Topic 3, additive strategy showed the largest coefficient estimate ($\beta_{3_{additive}} = 0.98$ with $p < .001$). Thus, additive strategy represented Topic 3 the most.

The results of the linear regression analysis are shown in Table 4. The coefficient of Topic 2 ($\theta_{topic_2} = 3.02$ with $p < .001$) was larger than the coefficient of Topic 1 ($\theta_{topic_1} = 2.75$ with $p < .001$). The was interpreted to mean that, when the teacher used Topic 2 more than Topic 1 to answer the item, the answer was likely to obtain a higher score. However, Topic 3 was not significant ($\theta_{topic_3} = 0.09$, with $p = 0.39$) for the scores.

4.2 Study 2: Results of Approach 2, sLDA and ANOVA

To select the best-fitting model, the values of Cao et al. (2009) and JSD indexes for two- to ten-topic models are shown in the right plot of Fig. 1. Because of the lowest

Cao et al. (2009) index (0.058) and the highest JSD index (2.369), the three-topic model was selected as the best-fitting model.

As with Study 1, we interpreted the three topics based on answers including discussions of mathematical operations and strategies. Topic 1 was labeled as Setting Up a Proportion Strategy. Representative written answers include the following:

*“You can set up a **proportion** to figure this out. It would be $(16/10) = (x/25)$. You cross-multiply, getting $10x = 16(25)$. $16(25)$ is 400 (since $25 \times 4 = 100$ and 4×4 is 16). So, your equation is now $10x = 400$. You **divide** by 10 and your answer is 40 cm.”*

*“40 cm high. You set up a **proportional** equation, **cross-multiply** and then solve for x . so $16/x = 10/25$ then **cross-multiply** to get $10(x) = 400$ and solve for x .”*

used the equations with multiplications, and the words expressing proportion strategy.

For Topic 2, we interpret it as Use of Multiplicative Comparison Strategies. Representative written answers for Topic 2 include the following:

*“The poster is 40 cm tall. The length was enlarged $2.5x$, so the **scale** would be the same for the height. 2.5×16 is 40.”*

*“The poster is 40 cm high. The **scale** is 1:2.5 which we can find out by looking at the width. We take that **scale** and **multiply** it by the height to find the poster height.”*

These all had use of words indicating multiplication equations and scale.

While Topic 3 was interpreted as Use of Additive Comparison Strategies. Representative answers for Topic 3 include the following:

*“The poster is 31cm high. I got my answer because I found the **difference** in width of the two posters ($25-10 = 15$). Then I took that solution of 15 and **added** it to the width of the first poster (16cm). I know that $15\text{cm} + 16\text{cm} = 31\text{ cm}$ for the width.”*

*“The poster is 31 cm high. The **difference** between 25 cm and 10 cm is 15. Therefore, I would **add** 15 cm to the original height of 16 cm.”*

*“The poster is 40 cm high. To get from 10 to 25, I doubled it and **added** half ($10+10+5$), so I did the same with the height. I doubled 16 to get 32 and **added** half to get 40.”*

These all contained the addition equations and words about additive strategy.

Results of the sLDA model are shown in Table 5. At the $\alpha = 0.05$ level, the estimated coefficient of Topic 2 ($\theta_{\text{topic}_2} = 3.12$ with $p < .001$) were larger than the estimated coefficients of Topic 1 and Topic 2 ($\theta_{\text{topic}_1} = 2.92$ and $\theta_{\text{topic}_3} = 0.46$ with $p < .001$). This indicated that using Topic 2 was more likely to obtain a higher score on the item.

Table 6 shows how strategies and topics were related. For Topic 1, the most significant factor was the proportion strategy ($\beta_{1_proportion} = 0.79$, $p < .001$). For Topic 2, the scale strategy showed the largest coefficient ($\beta_{2_scale} = 0.69$, $p < .001$). For Topic 3, additive strategy showed the largest coefficient ($\beta_{3_additive} = 0.88$, $p < .001$).

Table 5 Effects of problem-solving strategies on topics from sLDA

Topic	Est.	S.E.	<i>t</i> -stat.	<i>p</i> -value
Topic 1 Setting up a proportion strategy	3.00	0.09	34.90	<.001
Topic 2 Use of multiplicative comparison strategies	3.12	0.08	41.03	<.001
Topic 3 Use of additive comparison strategies	0.46	0.12	3.79	<.001

S.E. = standard error

Table 6 Effects of problem-solving strategies on topics from ANOVA

Strategy	Estimate	S.E.	<i>t</i> -statistic	<i>p</i> -value
Topic 1 Setting up a proportion strategy				
Additive	0.05	0.04	1.22	0.22
Proportion	0.79	0.03	29.23	<.001
Scale	0.22	0.02	10.86	<.001
Unit	0.32	0.06	5.32	<.001
Other	0.14	0.04	3.29	<.001
Topic 2 Use of multiplicative comparison strategies				
Additive	0.07	0.05	1.37	0.17
Proportion	0.65	0.02	28.27	<.001
Scale	0.69	0.05	14.78	<.001
Unit	0.42	0.07	6.23	<.001
Other	0.14	0.03	4.64	<.001
Topic 3 Use of additive comparison strategies				
Additive	0.88	0.03	29.53	<.001
Proportion	0.07	0.02	3.93	<.001
Scale	0.13	0.01	9.12	<.001
Unit	0.26	0.04	6.10	<.001
Other	0.17	0.03	5.83	<.001

S.E. = standard error

4.3 Comparison of Study 1 and Study 2

Table 7 showed the cosine similarity and JSD of per-document topic proportions estimated from STM in Study 1 and sLDA in Study 2. The cosine similarity values on the diagonal were close to 1, which indicated the pair of topics from STM and sLDA were similar. Whereas the off-diagonal values were close to 0, which meant the pair of topics were different. For JSD, the diagonal values were close to 0, which indicated that the pair of topics from STM and sLDA were not divergent to one another. While the off-diagonal values were close to 1, which showed the pair was divergent. Thus, the results of Study 1 and Study 2, were in agreement.

Table 7 Cosine similarity and JSD between results of STM in Study 1 and sLDA in Study 2

Cosine similarity		sLDA in Study 2		
		Topic 1	Topic 2	Topic 3
STM in Study 1	Topic 1	0.95	0.27	0.19
	Topic 2	0.43	0.94	0.35
	Topic 3	0.07	0.15	0.86
JSD		sLDA in Study 2		
		Topic 1	Topic 2	Topic 3
STM in Study 1	Topic 1	0.04	0.44	0.36
	Topic 2	0.44	0.07	0.51
	Topic 3	0.46	0.48	0.08

Note: Topic 1 = Setting up a proportion strategy; Topic 2 = Use of multiplicative comparison strategies; Topic 3 = Use of additive comparison strategies

5 Conclusion

In this study, we demonstrated the utility of topic modeling for detecting the relations among the examinees' written answers, problem-solving strategies, and scores. The results showed that it would be helpful to include multiple metadata into the model to improve the interpretability of the extracted topic structure underlying the collection of written materials. This study suggests that the inclusion of metadata could be a useful technique to analyze a collection of texts, especially for educational test datasets, regarding the examinees' thinking and reasoning procedures hidden in written answers. Even though the main purpose of this study was to show the applicability of both two-step approaches, the researcher would choose one of these approaches depending on their research questions, especially which relationship would be more important, either text-covariate or text-outcome. The findings of this study would be expanded as the research questions in the future. To provide the practical guidelines for the researchers, a simulation study manipulating the conditions of the data set, such as the number of written answers and the unique words, the average lengths of each written answer, or the types of covariates and the outcome variables, would be considered.

Funding This work was supported in part by the National Science Foundation under grants DRL-1751309 and DRL-1813760.

References

- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84.
- Blei, D. M., & McAuliffe, J. D. (2008). Supervised topic models. In *Advances in neural information processing systems* (pp. 121–128). Curran Associates.
- Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, 72(7), 1775–1781.

- Cardozo-Gaibisso, L., Kim, S., Buxton, C., & Cohen, A. (2020). Thinking beyond the score: Multidimensional analysis of student performance to inform the next generation of science assessments. *Journal of Research in Science Teaching*, 57(6), 856–878.
- Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, 17(1), 61–84.
- Kim, S., Kwak, M., Cardozo-Gaibisso, L., Buxton, C., & Cohen, A. S. (2017). Statistical and qualitative analyses of students' answers to a constructed response test of science inquiry knowledge. *Journal of Writing Analytics*, 1(1), 82–102.
- Mcauliffe, J., & Blei, D. (2007). Supervised topic models. *Advances in Neural Information Processing Systems*, 20.
- Nagwani, N. K. (2015). Summarizing large text collection using topic modeling and clustering based on MapReduce framework. *Journal of Big Data*, 2(1), 1–18.
- Paul, M., & Dredze, M. (2011). You are what you tweet: Analyzing Twitter for public health. *Proceedings of the Fifth International AAAI Conference on Web and Social Media*, 5(1), 265–272.
- Phan, X. H., Nguyen, L. M., & Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web* (pp. 91–100). Association for Computing Machinery.
- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4), 1064–1082.
- Schofield, A., Magnusson, M., & Mimno, D. (2017). Pulling out the stops: Rethinking stop-word removal for topic models. In *Proceedings of the 15th conference of European chapter of the Association for Computational Linguistics* (Vol. 2, pp. 432–436). Association for Computational Linguistics (ACL).
- Xiong, J., Choi, H. J., Kim, S., Kwak, M., & Cohen, A. S. (2019). Topic modeling of constructed-response answers on social study assessments. In *Annual meeting of the Psychometric Society* (Vol. 322, pp. 263–274). Springer.

Modeling Student's Response Time in an Attribute Balanced Cognitive Diagnostic Adaptive Testing



Tong Wu, Shaoyang Guo, and Hua-Hua Chang

Abstract Nowadays, more and more people are paying attention to processing data in different fields. Following the ease of computer implementation within the classroom and test environment, Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT) is a new advanced form of assessment. It combines the advantages of the Cognitive Diagnostic Model (CDM) and CAT, which could better assess students' learning acquirement. More and more universities are applying this technique to the university course assessment (Morphew et al, Phys Rev Phys Educ Res 14(2):020110, 2018). To improve the applicability, accompany modern processing data use wise, many researchers have started implementing Response Time (RT) into CD-CAT. Two methods implementing Response Time (RT) with fulfilling attribute-balancing constraints is proposed: Time Weighted Modified Maximum Global Discrimination Index and Time Weighted Modified Posteriori Weighted Kullback-Leibler Information. These methods are compared in simulation with time-weighted non-attribute balancing methods. The result shows the proposed methods provide researchers with precise measurement accuracy as well as improvement in test efficiency.

Keywords CD-CAT · Attribute-balancing · Response time · Measurement efficiency · Item exposure control

1 Introduction

Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT) (Chang, 2015; McGlohen & Chang, 2008; Xu et al., 2003) is a tailor-made test that provides fruitful information about students. Combining the advantages of CD and CAT,

T. Wu (✉) · H.-H. Chang
College of Education, Purdue University, West Lafayette, IN, USA
e-mail: wu473@purdue.edu

S. Guo
Institute of Curriculum and Instruction, Faculty of Education, East China Normal University, Shanghai, China

the personalized test assembly and accurate attribute estimation features make CD-CAT more and more popular especially in the pandemic season that students must take tests at home. However, avoid cheating at home has not yet been addressed by using psychometric methodology. It is urgent to ameliorate the current psychometric methodologies and apply those to practical usage. Rather than providing traditional summative scores in a test, RT often provides insights about students’ testing behavior, solution strategy, and cognitive demands of items. RT has been proven to be liable as an extra information in the item selection algorithms (Fan et al., 2012; Finkelman et al., 2014), attribute balance have not yet been considered which is indispensable in an university course assessment. This study extends the previous CD-CAT studies with attribute balances (Cheng, 2010) which makes the algorithm more applicable to the needs of university course assessment. The lognormal RT model (Van der Linden, 2007) is used for this study.

1.1 CD Models (CDM)

CDMs define examinees’ latent attribute through a vector $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_K)$, where K is the numbers of attributes. If the examinee mastered the attribute, then the specific attribute equals 1, otherwise, it would be 0. In order to select the appropriate items to measure examinees’ latent attribute, the item bank with items of latent attribute defined is called Q-matrix. Q-matrix represents the item bank in CDMs (Tatsuoka, 1985), where J is the total numbers of the item bank, and K is the attribute column. The q_{jk} represent the jth row and kth column entry of the Q-matrix, if $q_{jk} = 1$, then item j is measuring attribute k, and 0 vice versa. Figure 1 shows an example of Q-matrix, as an example, item 1 is measuring attribute 1, 3, 4:

Content experts determine the validation of Q-matrix prior to the tests (Cheng, 2009; Hsu et al., 2013) and it can also be adjusted and identified by empirical experiment (De La Torre, 2008).

The DINA Model Many CDMs have been proposed and developed over years. The model used in this study is the noisy input, deterministic input, noisy ‘and’ gate (DINA) model (Haertel, 1989; Junker & Sijtsma, 2001). It is a popular model which assumes all the skills are independent with each other. The model considers

Fig. 1 Example of Q-matrix

	q1	q2	q3	q4
Item 1	1	0	1	1
Item 2	0	1	0	0
Item 3	1	1	0	0
Item 4	1	0	0	1

two important behaviors in empirical testing: ‘guessing’ and ‘slipping’ parameters. ‘Slipping’ is a capable examinee makes a careless mistake by random reasons while ‘guessing’ is an incapable examinee makes a correct response by guessing. In order to check whether examinee is capable or not, the model includes an indicator shown as following:

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

Where $\eta_{ij} = 1$ represents the i th examinee is capable for the j th item, otherwise, i th examinee cannot answer j th item correctly, unless the examinee ‘guess’ it correctly. The DINA model then can be demonstrated as follows:

$$P(X_{ij} = 1 | \alpha_i) = (1 - s_j)^{\eta_{ij}} g_j^{1 - \eta_{ij}}. \tag{1}$$

The probability of answering the item correct is based on the examinee’s latent attributes and its production of ‘slipping’ probability with s_j to the power of the indicator and ‘guessing’ probability with g_j to the power of one minus the indicator. Both parameters and examinee’s attribute profile can be estimated by Maximum Likelihood Estimation (MLE). Maximum a- Posteriori (MAP) can be implemented to handle extreme cases for students’ attribute profile.

1.2 Item Selection in CD-CAT

Unlike CAT’s feature of measuring students’ latent traits, CD-CAT measures students’ discrete latent classes. Thus, instead of maximum Fisher information as item selection in CAT (Lord, 1980), the introduction of the Kullback-Leibler information (Chang & Ying, 1996) in CAT naturally satisfies the CD-CAT to measure discrete latent classes. The Kullback-Leibler information measures the ‘distance’ between two probability distributions $f(x)$ and $g(x)$ (Cover Thomas & Thomas Joy, 1991):

$$d[f, g] = E_f \left[\log \left[\frac{f(x)}{g(x)} \right] \right]$$

In CD, the Kullback-Liebler (KL) distance between two probability distributions, one is students’ response conditioning on the current latent attribute estimation, which is $f(X_{ij} | \hat{a}_i)$, and the other is students’ response conditioning on student’s true latent attributes, which is $f(X_{ij} | a_i)$, the formula of KL is shown as follow:

$$KL_j(\hat{a}_i | a_t) = \sum_{x=0}^1 \log \left[\frac{P(X_{ij} = x | \hat{a}_i)}{P(X_{ij} = x | a_t)} \right] P(X_{ij} = x | \hat{a}_i). \tag{2}$$

The KL index is measuring how deviate the distance is by answering the item between the true attribute and the estimated attribute. The larger the KL index, the more information we would obtain from the examinee for this item. While administering the items in the test, the true attribute is unknown and there are 2^k possible states for a certain examinee, thus, Xu et al. (2003) proposed the global discrimination index (GDI), which is equivalent to KL information given 2^k possible latent attributes:

$$GDI_j(\hat{a}_i) = \sum_{t=1}^{2^k} \left[\sum_{x=0}^1 \log \left[\frac{P(X_{ij} = x | \hat{a}_i)}{P(X_{ij} = x | a_t)} \right] P(X_{ij} = x | \hat{a}_i) \right] \tag{3}$$

The largest GDI selected from the item bank would be the next item to be administered to examinee. Based on GDI, Cheng (2009, 2010) considered the attributes balancing method (MMGDI) and the more powerful GDI implementing the posteriori distribution are so called Posteriori Weighted Kullback-Leibler Information Index (PWKL). The MMGDI method implement an attribute-balancing index (Cheng, 2010) would better suit the empirical test settings for attributes constraints and obtain smaller estimation error:

$$MMGDI_j(\hat{a}_i) = GDI_j(\hat{a}_i) \prod_{k=1}^K \left(\frac{B_k - b_k}{b_k} \right)^{q_{jk}}, \tag{4}$$

where B_k is the number of items required for k th attribute, and b_k is the item selected for k th attribute. q_{jk} indicates the specific item and attribute in the Q-matrix. Note that in CD assessment, an item normally contains more than one attribute, thus, it is reasonable to have the sum of all attributes number selected less than test length (L) prespecified:

$$\sum_{k=1}^K B_k \leq L.$$

Whenever the attribute-balancing index fulfills, the item selection would eliminate the attribute balance index and apply GDI only to finish up selecting items for the rest of the test.

The PWKL is a straightforward method that implements with posteriori distribution to GDI (Cheng, 2009):

$$PWKL_j(\hat{a}_i) = \sum_{t=1}^{2^k} \left\{ \pi(a_t | x_{n-1}) \left[\sum_{x=0}^1 \log \left[\frac{P(X_{ij} = x | \hat{a}_i)}{P(X_{ij} = x | a_t)} \right] P(X_{ij} = x | \hat{a}_i) \right] \right\} \tag{5}$$

Where $\pi(a_c | x_{n-1}) = p(a_c) \prod_{j=1}^{n-1} P(X_{ij} = 1 | a_c)^{x_{ij}} [1 - P(X_{ij} = 1 | a_c)]^{1-x_{ij}}$ is the posterior distribution and it’s been calculated based on the previous (n – 1) item responses from students. This method improves measurement precision compared to GDI especially in the situation of item banks containing large slipping and guessing parameters (Cheng, 2010).

1.3 Response Time Framework

Lognormal model is a popular model from its simple and practical prospect, the model is proposed by van der Linden (Van der Linden, 2007) shown as follow:

$$f(t_{ij} | \tau_i) = \frac{\alpha_j}{t_{ij} \sqrt{2\pi}} e^{-\frac{1}{2} [\alpha_j (\log t_{ij} - \beta_j + \tau_i)]^2}, \tag{6}$$

where t_{ij} represents the response time for examinee i on item j , τ_i is the latent speed parameter for examinee i , α_j and β_j are time discrimination parameters and time intensity parameters for j th item. The formula can be transferred into a normally distribution form, it is easily seen that $\mu_{ij} = \beta_j - \tau_i$ and $\sigma_j^2 = 1/\alpha_j^2$. Thus, for each item j , examinee i ’s response time given his/her speed ability, it follows the normal distribution as:

$$\log(t_{ij} | \tau_i) \sim N[\beta_j - \tau_i, 1/\alpha_j^2]. \tag{7}$$

The estimation methods for latent speed ability is using MLE and its likelihood function is shown below:

$$L(\tau_i) = \prod_{j=1}^J \frac{\alpha_j}{t_{ij} \sqrt{2\pi}} e^{-\frac{1}{2} [\alpha_j (\log t_{ij} - \beta_j + \tau_i)]^2}, \tag{8}$$

Since the true speed ability for each examinee is unknown, we need to substitute the MLE’s speed estimation by taking the first derivative as follow:

$$\hat{\tau}_i^{mle} = \frac{\sum_{j \in R_m} \alpha_j^2 (\beta_j - \log t_{ij})}{\sum_{j \in R_m} \alpha_j^2}. \tag{9}$$

The expected time to answer the j th item given the latent MLE estimation of speed is:

$$E [T_{ij} | \hat{\tau}_i^{mle}] = e^{(\beta_j - \hat{\tau}_i^{mle} + 1 / (2\alpha_j^2))}. \quad (10)$$

In order to increase measurement efficiency in RT aspect, Fan et al. (2012) demonstrated the MI combined response time (MIT) selection method:

$$j_{m+1} = \max_j \left\{ \frac{I_j(\hat{\theta}_i^{mle})}{E [T_{ij} | \hat{\tau}_i^{mle}]} : j \in R_m \right\}. \quad (11)$$

Using this formula, items contain high information will be chosen rather than the large time item. Fan et al. (2012) found the more skewness of item exposure for the MIT, thus they proposed the ASB with response time (ASBT) for balancing the item exposure:

$$j_{m+1} = \max_j \left\{ \frac{1}{E [T_{ij} | \hat{\tau}_i^{mle}] | b_j - \hat{\theta}_m |} : j \in R_m \right\}, \quad (12)$$

which shows balanced item exposure and more efficient test time duration.

2 Methods

This section first introduces the simulated data in this study and then the item selection algorithms. Next, four evaluation criteria are demonstrated.

2.1 Data

We simulated a 300-item bank with the noisy input, deterministic input, noisy ‘and’ gate (DINA) model (Haertel, 1989; Junker & Sijtsma, 2001), the guessing parameter $g \sim \text{unif}(0.05, 0.25)$, slipping parameter $s \sim \text{unif}(0.05, 0.25)$, time discrimination parameter $\alpha \sim \text{unif}(2, 4)$, time difficulty parameter $\beta \sim N(0, 0.25)$. For the Q-matrix simulation, we have 6 attributes and generate the Q-matrix entry by entry, each item should measure 20% of the attributes on average. 1000 examinees with speed parameters (τ) $\sim N(0, 1)$ are simulated. Each examinee’s attribute profile is randomly generated from $\text{unif}(0, 1)$. Test length of 30, minimum of 5 items are required to select from each attribute. Table 1 shows the numbers of items within

Table 1 Number of items measuring (or examining mastering) each attributes

	Attributes					
	A1	A2	A3	A4	A5	A6
Number of items	54	66	57	53	54	55
Number of examinees	477	508	507	499	483	519

Table 2 Number of items measuring (or examining mastering) each possible number of attributes

	0	1	2	3	4	5	6
Number of attributes	0	1	2	3	4	5	6
Number of items	0	199	80	16	5	0	0
Number of examinees	17	87	254	297	232	99	14

each attribute and the numbers of students acquire each attribute. The number of attributes for items and examinees are presented in Table 2.

Time Weighted Modified Maximum Global Discrimination Index (TW-MMGDI).

$$TW - MMGDI_j (\hat{\alpha}_i) = \frac{GDI (\hat{\alpha}_i)}{E [T_{ij} | \hat{\tau}_i^{mle}]} \prod_{k=1}^K \left(\frac{B_k - b_k}{b_k} \right)^{q_{jk}} \tag{13}$$

Time Weighted Modified Posteriori Weighted Kullback-Leibler Information (TW-MPWKL)

$$TW - MPWKL_j (\hat{\alpha}_i) = \frac{PWKL (\hat{\alpha}_i)}{E [T_{ij} | \hat{\tau}_i^{mle}]} \prod_{k=1}^K \left(\frac{B_k - b_k}{b_k} \right)^{q_{jk}} \tag{14}$$

Students’ achievement should be assessed by assigning enough items on each attribute for comprehensive estimation of knowledge acquirement and the completion time of each item to detect the speed. Thus, the proposed attribute balancing methods have an advantage over the non-attribute balancing methods. It can measure the students’ attribute profile more accurately by assigning specific numbers of each attribute of items within the test. On the other hand, the non-attribute balancing methods focus on selecting the deviation of the attribute profiles for item selection. Even though considering the ‘distance’ for selecting items is appropriate, the chosen items might be too few for some individual attribute estimation which makes the estimation inaccurate. Thus, results generated by non-attribute balancing item selection methods might mislead instructors’ interpretation of students’ knowledge acquirement. With the attribute constraints, the attribute balancing methods can measure each attribute comprehensively and improve the measurement accuracy for individual attribute and attribute profiles. Results generated by the attribute balancing methods would efficiently assist instructors’ interpretation and generate better guidance to students’ weakside of knowledge acquirement. Furthermore, the Response Time implementation improves the test efficiency, saving labor and money for operational test.

2.2 Evaluation Criteria

Four evaluation criteria are used to compare the methods:

1. Attribute Recovery Rate (ARR)

The Attribute Recovery Rate (ARR) is determined by the rate of true indication of attributes:

$$ARR_k = \frac{\sum_{i=1}^N A_{ik}}{N} = \frac{\sum_{i=1}^N (I_{(\hat{\alpha}_{ik}, a_{ik})})}{N}, (k = 1, 2, 3, \dots k)$$

2. Pattern Recovery Rate (PRR)

The Pattern Recovery Rate (PRR) is determined by the rate of true indication of attribute pattern:

$$PRR = \frac{\sum_{i=1}^N A_i}{N} = \frac{\sum_{i=1}^N (I_{(\hat{\alpha}_i, a_i)})}{N} \quad (15)$$

3. Average Time Unit

Average time unit is the mean test time duration recorded by the computer in the system.

4. The χ^2 statistic measures the skewness of item exposure rate distribution (Chang & Ying, 1999):

$$\chi^2 = \frac{\sum_{j=1}^J (r_j - L/J)^2}{L/J}, \quad (16)$$

where r_j represents the exposure rate of j th item, L represents the test length. J is the total number of items in the item pool. The smaller the χ^2 statistic is, the better the item exposure would be. There is no suggested criterion value for the statistics but smaller is better.

The variable-length χ^2 statistics is the following:

$$\chi^2 = \frac{\sum_{j=1}^J (r_j - \sum_{j=1}^J r_j / J)^2}{\sum_{j=1}^J r_j / J} \quad (17)$$

3 Results

The results compare item selection algorithm time-weighted attribute balanced methods with no time-weighted methods and no attribute balanced methods. Table 3 shows the simulation results for test length is 30. By taking advantage of accurate measurement in MMGDI and considering RT, TW-MMGDI comprehensively improve from three aspects: decreasing time of test (from 56.9 to 54.6, about 4% decrease), lowering item exposure statistic (from 156.8 to 133.8, about 14.7% decrease) and maintaining high PRR compared to the no time-weighted method. TW-GDI has a relatively shorter test time and item exposure statistic compared to MMGDI. It is mainly because no attribute balance constraints were applied, and RT is implemented in the algorithm. TW-MMGDI outperforms TW-GDI in measurement accuracy as well as test time efficiency. TW-MPWKL also decreases the period of test (from 56.9 to 54.9, about 3.5% decrease), item exposure statistic (from 85.7 to 79.6, about 7.1% decrease) and maintains high PRR compared to MPWKL. Turning to the comparison of the two proposed time-weighted attribute balancing methods, TW-MPWKL obtains a better item exposure statistic than TW-MMGDI without any item exposure control methods implemented.

Table 4 shows the simulation results for test length is 40. TW-MMGDI outperforms with decreasing time of test (from 76.7 to 73.99, about 3.5% decrease), item exposure statistic (from 97.6 to 87.3, about 11% decrease) and maintains high PRR compared to the no time weighted method. As the test length increase, the improvement in measurement criteria decreases.

The results show the RT implemented item selection algorithms is feasible in practical settings of relatively low stakes test that requires high measurement accuracy. These methods not only shorten the test time 3–4% but also improve item exposure statistics up to 7% for home testing while maintaining high PRR, ARR.

Table 3 Results for simulation study L = 30

Method	A1	A2	A3	A4	A5	A6	Time	χ^2	PRR
TW-MMGDI	.99	.99	.99	.99	.99	.98	54.60	133.8	.968
MMGDI	.99	.99	.99	.99	.99	.99	56.92	156.8	.960
TW-GDI	.94	.94	.94	.95	.95	.96	55.49	18.2	.719
TW-MPWKL	.99	.99	.99	.99	.99	.99	54.96	79.6	.959
MPWKL	.99	.99	.99	.99	.99	.99	56.89	85.7	.965
TW-PWKL	.91	.92	.94	.94	.94	.96	54.25	46.8	.689

Note: *TW-MMGDI* Time Weighted Modified Maximum Global Discrimination Index, *MMGDI* Modified Maximum Global Discrimination Index, *TW-GDI* Time Weighted Global Discrimination Index, *TW-MPWKL* Time Weighted Modified Posteriori Weighted Kullback-Leibler Information, *MPWKL* Modified Posteriori Weighted Kullback-Leibler Information, *TW-PWKL* Time Weighted Posteriori Weighted Kullback-Leibler Information, TW-GDI and TW-PWKL do not consider attribute constraint

Table 4 Results for simulation study L = 40

Method	A1	A2	A3	A4	A5	A6	Time	χ^2	PRR
TW-MMGDI	.99	.98	.99	.99	.99	.99	73.99	87.3	.945
MMGDI	.99	.99	.99	.99	.99	.99	76.68	97.6	.942
TW-GDI	.95	.94	.95	.95	.95	.95	72.68	31.2	.740
TW-MPWKL	.99	.98	.99	.98	.99	.99	73.44	66.5	.961
MPWKL	.99	.99	.99	.99	.99	.99	75.68	66.9	.966
TW-PWKL	.94	.93	.95	.95	.96	.96	72.33	49.2	.732

Note: *TW-MMGDI* Time Weighted Modified Maximum Global Discrimination Index, *MMGDI* Modified Maximum Global Discrimination Index, *TW-GDI* Time Weighted Global Discrimination Index, *TW-MPWKL* Time Weighted Modified Posteriori Weighted Kullback-Leibler Information, *MPWKL* Modified Posteriori Weighted Kullback-Leibler Information, *TW-PWKL* Time Weighted Posteriori Weighted Kullback-Leibler Information, TW-GDI and TW-PWKL do not consider attribute constraint

Table 5 Results for variable length simulation (Termination rule = (0.8, 0.1))

Method	A1	A2	A3	A4	A5	A6	Time	χ^2	PRR	Test length
TW-MMGDI	.98	.99	.99	.99	.98	.99	50.35	163.01	.945	27.89
MMGDI	.99	.99	.99	.99	.99	.99	53.05	182.25	.959	27.76
TW-GDI	.95	.94	.96	.95	.96	.95	89.96	34.30	.757	49.34
TW-MPWKL	.99	.99	.99	.99	.99	.99	51.15	72.46	.948	28.32
MPWKL	.99	.99	.99	.99	.99	.99	53.58	101.33	.953	28.29
TW-PWKL	.95	.94	.94	.94	.95	.96	90.23	50.38	.763	49.31

Note: *TW-MMGDI* Time Weighted Modified Maximum Global Discrimination Index, *MMGDI* Modified Maximum Global Discrimination Index, *TW-GDI* Time Weighted Global Discrimination Index, *TW-MPWKL* Time Weighted Modified Posteriori Weighted Kullback-Leibler Information, *MPWKL* Modified Posteriori Weighted Kullback-Leibler Information, *TW-PWKL* Time Weighted Posteriori Weighted Kullback-Leibler Information, TW-GDI and TW-PWKL do not consider attribute constraint

The methods are also extended into variable length, the preliminary results are shown in Table 5 with termination rules of the largest attribute pattern greater than 0.8 and the second largest attribute pattern less than 0.1 (Hsu et al., 2013). The other conditions remain the same as previous. Comparing TW-MMGDI and MMGDI, the time implemented method improves item exposure and time simultaneously while maintaining comparable PRR measurement accuracy. TW-MPWKL and MPWKL show similar results. Comparing both time-weighted methods, TW-MPWKL has a natural advantages of item exposure and it obtains the features of maintaining the measurement accuracy as well as improve test efficiency with shorter test time. The methods are practically useful to implement in university courses and utilize the item bank wisely.

4 Discussion

The research aims to explore the applicability of the implementation of RT and attribute balancing in CD-CAT settings. Results indicated that the proposed methods could enhance Test time, item exposure, and maintain comparable measurement accuracy. The improvement is better when the test length is short. Especially in a classroom setting where students could be disengaged easily, the test within a short period could gather more information with accurate measurement would be preferred.

Limitations exist within this study. First, an item bank should normally be at least 12 times as many items as the test length (Stocking, 1994). This study contains a relatively small item bank which is in a similar condition to a university course. A larger item bank would lead to a more significant difference between methods. Secondly, under the fixed-length situation, the time-weighted attribute balancing methods are proved to be more efficient. The variable-length situation needs more investigation under various test lengths and times. The information is gathered from statistical simulations, an empirical study should be investigated to validate the conclusion in future research. Lastly, this study utilizes the DINA model to demonstrate the simulation in CD-CAT. It can be easily extended to the different CDM and RT models for validation. The current study only considers CDMs and RT estimation and item selections separately. Researchers should consider joint modeling estimations and item selection algorithms of CDMs and RT models in the future, which might require more extensive computation power but improve more in the efficiency of information gathering in CD-CAT test settings.

Appendix

```
### This is the Item bank generation Code
s = runif(300,0.05,0.25)
g = runif(300,0.05,0.25)
Qmatrix = matrix(0,300,6)
## If 300 items and 4 attributes then we need to have 30%
  to be compared, if 6 attributes, then 20%
for (i in 1:300){
  for (j in 1:6){
    k = runif(1,0,1)
    if(k < 0.2){
      Qmatrix[i,j] = 1
    }
    else{Qmatrix[i,j]=0}
  }
}
alpha = runif(300,2,4)
beta = runif(300,0,.25)
Qmatrix = cbind(s,g,Qmatrix,alpha,beta)
```

```

Item_num = seq(1,300)
num_select1 = rep(0,300)
Qmatrix = cbind(Item_num,Qmatrix, num_select1)
colnames(Qmatrix) = c('Item_ID', 's', 'g', "q1", "q2", "q3", "q4",
  "q5", "q6", 'alpha', 'beta', 'num_select')
Itembank = as.data.frame(Qmatrix )
Examinee_attribute = matrix(0,1000,6)
  for (i in 1:1000){
    for (j in 1:6){
      k = runif(1,0,1)
      if(k < 0.5){
        Examinee_attribute[i,j] = 1
      }
      else{Examinee_attribute[i,j]=0}
    }
  }
##Examinee's RT parameter
tau = rnorm(N_examinee,0,1)
Examinee = as.data.frame(cbind(seq(1,N_examinee), Examinee_
  attribute,tau))
colnames(Examinee) = c('Examinee', 'A1', 'A2', 'A3', 'A4', 'A5',
  'A6', 'tau' )
## This is the TWMMGDI Method Code
TWMMGDI = function(valid_Itembank,ContentControl,MLE_nodes,est_
  attri,est_tau,Examinee_test){
  vector_ALPHA=valid_Itembank$alpha
  vector_BETA=valid_Itembank$beta
  vector_S=valid_Itembank$s
  vector_G=valid_Itembank$g
  q1 = valid_Itembank$q1
  q2 = valid_Itembank$q2
  q3 = valid_Itembank$q3
  q4 = valid_Itembank$q4
  q5 = valid_Itembank$q5
  q6 = valid_Itembank$q6
  n_item_Examinee_test=sum(is.na(Examinee_test[,1])==FALSE)
  current_Examinee_test=Examinee_test[1:n_item_Examinee_test,]
  vector_resp=current_Examinee_test$RESP
  responded_s = current_Examinee_test$s
  responded_g = current_Examinee_test$g
  q11 = current_Examinee_test$q1
  q21 = current_Examinee_test$q2
  q31 = current_Examinee_test$q3
  q41 = current_Examinee_test$q4
  q51 = current_Examinee_test$q5
  q61 = current_Examinee_test$q6
  lamda_est=(est_attri [1]^q1)*(est_attri [2]^q2)*(est_attri [3]^q3)
  *(est_attri [4]^q4)*(est_attri [5]^q5)*(est_attri [6]^q6)
  P_est=Dina(vector_S,vector_G,lamda_est)
  lamda_nodes=matrix(NA,nrow(valid_Itembank), 64)
  P_nodes=matrix(NA,nrow(valid_Itembank), 64)
  gdi0=matrix(NA,nrow(valid_Itembank), 64)
  gdi1=matrix(NA,nrow(valid_Itembank), 64)
  posteriori_weighted = matrix(NA,length(vector_resp), 64)

```

```

P_responded = matrix(NA,length(vector_resp),64)
lamda_responded = matrix(NA,length(vector_resp),64)
for (i in 1:64){
lamda_nodes[,i]=(MLE_nodes[i,1]^q1)*(MLE_nodes[i,2]^q2)*
(MLE_nodes[i,3]^q3)*(MLE_nodes[i,4]^q4)*(MLE_nodes[i,5]^q5)*
(MLE_nodes[i,6]^q6)
  P_nodes[,i]=Dina(vector_S,vector_G,lamda_nodes[,i])
  lamda_responded[,i] = (MLE_nodes[i,1]^q11)*
(MLE_nodes[i,2]^q21)*(MLE_nodes[i,3]^q31)*(MLE_nodes[i,4]^q41)*
(MLE_nodes[i,5]^q51)*(MLE_nodes[i,6]^q61)
  P_responded[,i] = Dina(responded_s,responded_g,lamda_
responded[,i])
  gdi0[,i]=(1-P_est)*log((1-P_est)/(1-P_nodes[,i]))
  gdi1[,i]=P_est*log(P_est/P_nodes[,i])
  for (j in 1:length(vector_resp)){
    posteriori_weighted[j,i] = P_responded[j,i]^vector_resp[j]*
(1-P_responded[j,i])^(1-vector_resp[j])
  }
}
pai=rep(NA,64)
for(i in 1:64){
  pai[i] = (1/2)^6*prod(posteriori_weighted[,i])
}
GDI=apply((gdi0+gdi1),1,sum)
BkI=apply(t(((ContentControl$Low-ContentControl$CUR)/
ContentControl$Low)^t(valid_Itembank[,4:9])),1,prod)
ET=exp(vector_BETA-est_tau+(1/(2*(vector_ALPHA^2))))
TWMGDIIndex=GDI*BkI/ET
valid_Itembank$TWMGDIIndex=TWMGDIIndex
return(list(valid_Itembank))
}

```

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. <https://doi.org/10.1177/014662169602000303>
- Chang, H.-H., & Ying, Z. (1999). Alpha-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23(3), 211–222. <https://doi.org/10.1177/01466219922031338>
- Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, 74(4), 619.
- Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: The modified maximum global discrimination index method. *Educational and Psychological Measurement*, 70(6), 902–913.
- Cover Thomas, M., & Thomas Joy, A. (1991). *Elements of information theory* (3rd ed., pp. 37–38). Wiley.
- De La Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, 45(4), 343–362.

- 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. <https://doi.org/10.3102/1076998611422912>
- Finkelman, M., Kim, W., Weissman, A., & Cook, R. (2014). Cognitive diagnostic models and computerized adaptive testing: Two new item-selection methods that incorporate response times. *Journal of Computerized Adaptive Testing, 2*(3), 59–76.
- 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.
- Hsu, C.-L., Wang, W.-C., & Chen, S.-Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement, 37*(7), 563–582.
- 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.
- Lord, F. M. (1980). *Practical applications of item response theory*. Lawrence Erlbaum Associates.
- McGlohen, M., & Chang, H.-H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods, 40*(3), 808–821.
- Morphew, J. W., Mestre, J. P., Kang, H. A., Chang, H. H., & Fabry, G. (2018). Using computer adaptive testing to assess physics proficiency and improve exam performance in an introductory physics course. *Physical Review Physics Education Research, 14*(2), 020110.
- Stocking, M. L. (1994). Three practical issues for modern adaptive testing item pools 1. *ETS Research Report Series, 1994*(1), i–34.
- Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics, 10*(1), 55–73.
- Van der Linden, W. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287. <https://doi.org/10.1007/s11336-006-1478-z>
- Xu, X., Chang, H., & Douglas, J. (2003). *Computerized adaptive testing strategies for cognitive diagnosis*. Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Canada

Impact of Construct Reliability on Proposed Measures of Structural Fit When Detecting Group Differences: A Monte Carlo Examination



Graham G. Rifenbark

Abstract Structural fit indices (SFIs) have been advanced due to the influence of the measurement model on the global fit indices (GFIs). First, GFIs are overly weighted by the measurement model. Second, GFI cut-offs were not determined in the context of varying magnitudes of the factor loadings; as a result, model fit seems to improve as the magnitude decreases, known as the *reliability paradox*. The focus of this study was to examine the relative performance of the recently proposed SFIs in their ability to detect a misspecified mean structure or covariance structure. This study was executed in the context of multiple group models where the misspecifications were in the form of true differences between populations. Of key interest was the impact construct reliability had on power rates for these SFIs, as well as how they performed relative to GFIs. Findings show that structural measures of fit outperformed the global measures of fit regardless of the type of misfit (e.g., mean or covariance). Measures of fit were more sensitive to the magnitude of the factor loadings when the covariance structure was misspecified, relative to when the mean structure was misspecified.

Keywords Structural equation modeling · Construct reliability · Multiple group · Structural misspecification · Statistical power · Goodness-of-fit

1 Introduction

Recently, structural fit indices (SFI) have been developed with the aim of evaluating approximate fit of the structural model in isolation. Specifically, Lance et al. (2016) developed a suite of SFIs based on conditions 9 and 10 by James et al. (1982) and

G. G. Rifenbark (✉)
University of Connecticut, Storrs, CT, USA
e-mail: graham.rifenbark@uconn.edu

earlier, McDonald and Ho (2002) developed a structural version of the root mean square error of approximation (RMSEA)—referred to as RMSEA-Path (RMSEA-P). The motivation for the development of these SFIs was based on the measurement model possessing a large proportion of the global degrees of freedom, therefore, a well fitting measurement model could mask a poor fitting structural model (Lance et al., 2016; McDonald & Ho, 2002) when utilizing conventional global fit indices (GFIs) and their common cut-offs per Hu and Bentler (1998, 1999).

To date, research has been conducted to evaluate the performance of GFIs while varying aspects related to construct reliability, however, this space has not been explored when evaluating SFI. With respect to construct reliability, key model characteristics include both the magnitude of factor loadings and the number of indicators per factor (Gagne & Hancock, 2006). Factor loadings convey the amount of variance explained by the latent variable, representing common variance and therefore, factor loadings that are large in magnitude are more easily able to discriminate between those who are high and low on the measured construct. The number of indicators per factor is also important seeing as factors with more indicators are said to be more reliably measured than those with fewer indicators holding the magnitude of the factor loadings constant (Gagne & Hancock, 2006).

An early example of research that considers characteristics of construct reliability and model fit was done by Fornell and Larcker (1981). Specifically, they introduced a method of estimating construct reliability and discuss the interplay between measurement and theory. To illuminate what constitutes *measurement* and *theory* it is easiest to imagine an observed variance-covariance matrix that contains observed variables that are thought to measure multiple constructs. Measurement is taken as elements that correspond to a group of observed variables that measure the same construct, therefore, if commonalities are high among these observed variables this would translate to a well defined construct (e.g., the magnitude of the factor loadings would be large). On the other hand, theory corresponds to relations among observed variables that are hypothesized to belong to different constructs. Therefore, if these elements in the observed variance-covariance matrix go against theory (e.g., covariances are positive when the hypothesized constructs are thought to be negatively related) then this would constitute poor theory. Fornell and Larcker (1981) found that as measurement decreases, the χ^2 test statistic improves (e.g., decreases and the null hypothesis is accepted); meanwhile, as theory decreases, the χ^2 test statistic also decreases (indicating good fit). In sum, Fornell and Larcker (1981) state that proper evaluation of the structural model can be misguided due to properties of the measurement model and therefore, consulting the chi-square test of model fit may not be appropriate when testing theory and propose means for estimating construct reliability.

More recent research has been carried out with respect to GFIs. Specifically, Kenny et al. (2015) show that the RMSEA tends to reflect an acceptable fitting model when there are a large number of degrees of freedom whereas, the CFI and TLI tend to reflect the opposite (Ding et al., 1995). On the other hand, Hancock and Mueller (2011) shed light on a phenomenon they call the *reliability paradox* via a population analysis. Specifically, given an identical structural misspecification, as

the magnitude of factor loadings decrease, model fit appears to improve, whereas, as the magnitude of factor loadings increase, model fit appears to worsen when consulting Hu and Bentler cut-offs for GFIs. As a solution, Hancock and Mueller (2011) proposed structural analogs of common GFIs and claim that their procedure, which requires two stages of estimation, removes the impact of the measurement model (i.e., the magnitude of factor loadings) and provides a reasonable evaluation of structural model fit. However, in a recent study, Rifenbark (2019) illustrated that the approach for constructing SFIs given by Hancock and Mueller (2011) does not accomplish what it purports to.

The purpose of this study was to determine the impact of construct reliability on the RMSEA-P and the conditions 9 and 10 (C9 and C10) SFIs when attempting to detect structural misspecifications in either the covariance or mean structures. To date, neither the RMSEA-Path, C9 or C10 SFIs have been evaluated while systematically varying facets of construct reliability when the covariance structure was misspecified nor have these SFIs been evaluated in the context of a misspecified mean structure. The evaluation of the structural component of a latent variable model can take on many different forms depending on the context. One such context that is widely used in the social and educational sciences are multiple group structural equation models [MG-SEM; Sörbom (1974)] and therefore, was a natural choice to accomplish the aims of this study.

Prior to formally presenting and reporting on the Monte Carlo simulation executed for this study, a brief introduction to latent variable models will be given along with how data-model fit is judged, as well as key differences between models (e.g., CFA and SEM) will be exposed, and finally, the methods available for evaluating structural model fit will be formally introduced and discussed.

2 Latent Variable Modeling

In the social and educational sciences, researchers routinely are unable to observe phenomenon directly and therefore, utilize latent variable models to measure these unobservable phenomenon. Latent variable models afford the opportunity to account for measurement error allowing inferences regarding structural relations (i.e., among measured constructs) to be made error-free. As an initial step, a confirmatory factor analysis (CFA) is executed and can include both mean and covariance structures. Equations for the model-implied covariance matrix ($\hat{\Sigma}$) and model-implied mean vector ($\hat{\mu}$) are presented below.

$$\begin{aligned} \hat{\Sigma} &= \Lambda \Phi \Lambda^t + \Theta \\ \hat{\mu} &= \tau + \Lambda \alpha \end{aligned} \tag{1}$$

In Eq. (1) with respect to the covariance structure, Λ is a pattern matrix and contains factor loading estimates that represent common variance among observed

variables that load onto the same factor. Φ is a symmetric matrix that corresponds to the latent variance-covariance matrix; specifically, the diagonal elements of Φ are the latent variances in the population and the off-diagonal elements of Φ correspond to the relations among latent variables (e.g., bi-directional paths). Finally, Θ , corresponds to measurement error (i.e., variance that is unrelated to the measured construct). With respect to the mean structure, α is a vector of latent means; τ is a vector of intercepts for the observed variables (i.e., expected value when the latent mean equals zero); and Λ is as before. Once a CFA model is found to be acceptable, structural equation modeling (SEM) is typically utilized to test theory. The key difference between CFA and SEM then, is that based on theory certain paths among latent variables are fixed to zero, while other paths are changed such that they are uni-directional (e.g., causal relations). In other words, endogenous latent variables are regressed onto exogenous variables and therefore, the system of equations to model need to be augmented (Widaman & Thompson, 2003):

$$\hat{\Sigma} = [(\hat{\tau} + \hat{\Lambda}(\mathbf{I} - \hat{\mathbf{B}})^{-1}\hat{\alpha})(\hat{\tau} + \hat{\Lambda}(\mathbf{I} - \hat{\mathbf{B}})^{-1}\hat{\alpha})' + [\hat{\Lambda}(\mathbf{I} - \hat{\mathbf{B}})^{-1}\hat{\Psi}(\mathbf{I} - \hat{\mathbf{B}})^{-1}\hat{\Lambda}' + \hat{\Theta}] \quad (2)$$

Looking at Eq. (2), $\hat{\Sigma}$ now depends on \mathbf{I} , $\hat{\mathbf{B}}$, and $\hat{\Psi}$. \mathbf{I} is an identity matrix with as many rows and columns as there are latent variables. Latent regression coefficients are stored in \mathbf{B} which has the same order as \mathbf{I} . Diagonal elements of \mathbf{B} must be zero, whereas, elements below the diagonal contain the latent regression parameters that can be freely estimated. Finally, Ψ is a matrix with the same dimensions as \mathbf{B} . The diagonal elements of Ψ correspond to the latent variances (exogenous variables) or the latent disturbances (endogenous variables); whereas, the off-diagonal elements correspond to covariances between the latent disturbances.

$$\hat{\mu} = \hat{\tau} + \hat{\Lambda}(\mathbf{I} - \hat{\mathbf{B}})^{-1}\hat{\alpha}. \quad (3)$$

With respect to the structural model, looking at Eq. (3), $\hat{\mu}$ now depends on \mathbf{I} and $\hat{\mathbf{B}}$ (defined above) and $\hat{\alpha}$ is the same as in Eq. (1); further, all measurement model matrices remain the same as before.

2.1 Model Fit

Model fit for latent variable models, regardless of whether a CFA or SEM is estimated, is determined by the model's ability to reproduce the observed variance-covariance matrix and mean vector. The task at hand for model estimation is to determine the set of parameter estimates (e.g., maximum likelihood estimates) that minimize the difference between the model-implied and observed moments and is referred to as the discrepancy fit function, F_{ML} . Considering the mean

and covariance structures simultaneously, Browne and Arminger (1995) write the discrepancy fit function as:

$$F_{ML}(\hat{\Sigma}, \Sigma; \hat{\mu}, \mu) = (\mu - \hat{\mu})^t \hat{\Sigma}^{-1} (\mu - \hat{\mu}) + \ln |\Sigma| + \text{tr}(\Sigma \hat{\Sigma}^{-1}) - \ln |\hat{\Sigma}| - q, \quad (4)$$

where q corresponds to the total number of free parameters across both the measurement and structural models. Using F_{ML} , a likelihood test statistic (T_{ML}) can be computed: $T_{ML} = F_{ML} * N$, where N corresponds to sample size. Assuming multivariate normality, T_{ML} is χ^2 distributed with its degrees of freedom equal to $\frac{P(P+3)}{2} - q$, where P corresponds to the number of observed variables. χ^2 is a test of exact model fit, however, due to its reliance on sample size this fit statistic can be sensitive to minor misspecifications, given a large sample size. It is important to note that when evaluating the fit of a CFA model, misfit from the structural model ($\hat{\Phi}$ or $\hat{\alpha}$) is not possible. This is because all latent parameters, across both mean and covariance structures are freely estimated. However, when moving to SEM, elements in \hat{B} and $\hat{\Psi}$ are fixed to zero and therefore, degrees of freedom are gained. This increase in the degrees of freedom affords the opportunity for misfit to stem from the structural model.

3 Approaches for Evaluating Structural Fit

James et al. (1982) developed conditions 9 and 10 to assess whether causal relations among latent variables were correctly specified. Condition 9 is satisfied when a hypothesized non-zero relationship between LVs that is confirmed to be non-zero in the population. For instance, if a given element in \hat{B} is found to be significantly different from zero. On the other hand, condition 10 is satisfied by nested χ^2 test that confirm that a hypothesized null relationship between LVs, is in fact null in the population. It is no surprise then, that the standard approach for evaluating structural model fit relies on nested chi-square tests ($\Delta\chi^2$) and relies on the estimation of two models: the correlated factors model (or CFA) and the hypothesized structural model (or SEM). The correlated factors model (or CFA) estimates all relations among latent variables (i.e., the off-diagonal of ϕ) via bi-directional paths, whereas the hypothesized SEM utilizes uni-directional paths and fixes certain paths to zero. Therefore, to evaluate the fit of the hypothesized SEM the difference in the test statistic (χ^2) between the CFA and SEM is compared to the critical value which is a function of the change in degrees of freedom between the two models.

3.1 RMSEA-Path

McDonald and Ho (2002) developed the RMSEA-P to gauge approximate fit of the structural model. Similar to the global RMSEA, the RMSEA-P provides the

amount of misfit per degree of freedom in the structural model. In order to construct the RMSEA-P, information from both the CFA and the hypothesized SEM are used to construct the structural analog of McDonald's d : $d_{path} = \frac{(\Delta\chi^2 - \Delta df)}{(N-1)}$, where N corresponds to the sample size.

3.2 Conditions 9 and 10

Lance et al. (2016) developed a suite of SFIs that were either condition 9 (C9) or condition 10 (C10) using one-of-three approaches: χ^2 , non-centrality [$\chi^2 - df$], or ratio [$\frac{\chi^2}{df}$] based. In order to construct the C9 and C10 SFIs, it is necessary to estimate the CFA, hypothesized SEM, and a null structural model. The null structural model is one in which all relations among the hypothesized exogenous and endogenous latent variables are fixed to zero. Therefore, the fit of the structural model is believed to fall somewhere on the continuum between the worst fitting structural model (the null structural model) and the best fitting structural model (the CFA, due to all possible latent parameters being freely estimated), therefore, regardless of whether a C9 or a C10 SFI is constructed, the denominator is the same—see Eq. (5) for both the C9 and C10 indices using the non-centrality approach:

$$\begin{aligned} C9 &= \frac{(\chi_{null}^2 - \chi_{SEM}^2) - (df_{null} - df_{SEM})}{(\chi_{null}^2 - \chi_{CFA}^2) - (df_{null} - df_{CFA})}; \\ C10 &= \frac{(\chi_{SEM}^2 - \chi_{CFA}^2) - (df_{SEM} - df_{CFA})}{(\chi_{null}^2 - \chi_{CFA}^2) - (df_{null} - df_{CFA})} \end{aligned} \quad (5)$$

Ultimately, Lance et al. (2016) recommend the use of the latter two approaches and recommend values of 0.99 or greater for C9 SFIs and 0.01 or less for C10 SFIs as support for an acceptable fitting structural model.

4 Simulation Study

4.1 Method

To determine the impact of construct reliability on the performance of the selected structural measures of fit, the following factors were manipulated: magnitude of factor loadings, number of indicators per factor, and group sample sizes. In terms of group differences, population differences were either small, medium, or large and were generated in either the mean or the covariance structure.

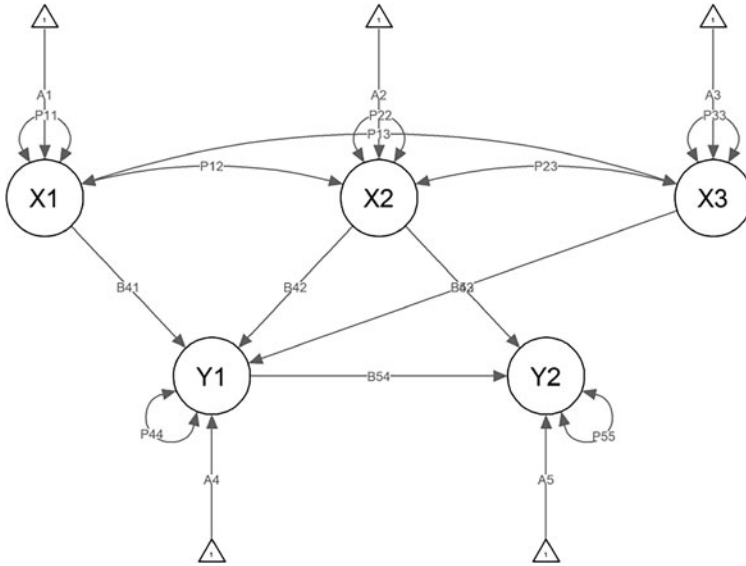


Fig. 1 Path diagram: Data Generating Model. Single group model with no misspecifications. Exogenous latent variables: X1, X2, X3. Endogenous latent variables: Y1 and Y2

Data Generation The data generating model was motivated by MacCallum (1986) and is depicted in Fig. 1. The plot was generated in R (R Core Team, 2017) using the semPlot package (Epskamp and with contributions from Simon Stuber, 2017) and for clarity the measurement model is omitted. The model contains five latent variables: three are exogenous (X1, X2, and X3) and 2 are endogenous (Y1 and Y2). Y1 is regressed onto each exogenous latent variable, whereas, Y2 is regressed onto the latent variables Y1 and X2. Data were generated from the multivariate normal distribution and the total variance for both the observed and latent variables were fixed to 1.0. Due to the focus on structural misspecification, full measurement invariance was generated between the two groups (i.e., all measurement parameters were identical across groups). All data were generated in R using the simsem package (Pornprasertmanit et al., 2016).

Manipulated Factors The magnitude of the factor loadings varied between 0.4, 0.6, and 0.8 which translates to indicator reliabilities of 0.16, 0.36, and 0.64, respectively. The number of indicators per factor (p:f) varied between 3 and 5 to determine the impact model size had on the performance of the structural measures of fit. When standardized factor loadings were 0.4 in the population, construct reliability was 0.36 (p:f = 3) or 0.49 (p:f = 5); when standardized factor loadings were 0.6, construct reliability was 0.63 (p:f = 3) or 0.74 (p:f = 5); and when standardized factor loadings were 0.8, construct reliability was 0.84 (p:f = 3) or 0.90 (p:f = 5)—it is important to note that due to the tau-equivalence among factor loadings, construct reliability equals coefficient H (Gagne & Hancock, 2006).

Sample size was fixed to 2,000, however, group sample sizes were manipulated. Specifically, the reference group sample size was either: 600, 1000, or 1400 with the focal group making up the difference (e.g., when the reference group contained 600 cases the focal group contained 1400 cases).

Group differences were generated either in the mean structure or the covariance structure. The magnitude of the generated differences spanned three levels. When group differences existed in the mean structure, the latent mean for X3 was fixed to 0.0 for the reference group, while the mean for X3 varied between 0.20 (small effect, $d = 0.2$), 0.50 (medium effect, $d = 0.5$), or 0.80 (large effect, $d = 0.8$) for the focal group—these effects mimic that of prior research (Fan & Sivo, 2009). When population differences were present in the covariance structure, the structural path emitting from X2 to Y2 was selected. Specifically, the standardized regression path was 0.30 for the reference group and 0.50 for the focal group (small effect, $d = .2$), 0.20 for the reference group and 0.60 for the focal group (medium effect, $d = 0.4$), or 0.00 for the reference group and 0.60 for the focal group (large effect, $d = 0.6$)—these were chosen based on prior research (Kang et al., 2016).

Group difference conditions were crossed with the other manipulated factors, therefore, when evaluating group differences a total of 54 [$3 \times 2 \times 3 \times 3$] unique conditions were investigated for both the mean structure and the covariance structure. Thus, in sum, a total of 108 simulation conditions were investigated with 1000 replications executed per condition.

Estimated Models The models required to be estimated were the null structural model, the hypothesized SEM, and the CFA. For each of these models, full metric and scalar invariance was modeled (i.e., factor loadings and manifest intercepts were constrained to be the same across groups). The null structural model estimates the latent covariances among all exogenous latent variables (X1, X2, and X3) and were not constrained to be the same across groups; on the other hand, all paths between the exogenous and the hypothesized endogenous latent variables (Y1 and Y2) were fixed to zero for both groups. With respect to the latent variances and means, these were freely estimated across groups. The hypothesized SEM was the data generating model, therefore, all of the correct structural paths were estimated (e.g., X1 to Y2 was fixed to zero) and constrained to be the same across groups. By doing so, full structural invariance was modeled and as a result, structural misspecification was introduced. The CFA model estimates all possible latent parameters (e.g., covariances, variances, and means) without any constraints across groups.

To construct RMSEA-P and $\Delta\chi^2$, information from the hypothesized SEM and CFA were utilized, while information from all three models was required to construct the C9 and C10 SFIs. All models were fitted in R using the lavaan package (Rosseel, 2012) with the sem function, invoked the mimic = "mplus" option and employed maximum likelihood estimation. Model identification and scaling was accomplished using the marker variable method whereby the factor loading and the manifest intercept of the first indicator for each latent variable were fixed to 1.0 and 0.0, respectively.

Outcomes To understand the impact construct reliability had on measures of structural and global model fit, descriptive statistics were estimated separately from replications that had either a misspecification in the mean or covariance structure. To quantify the performance of the various measures of fit, statistical power was determined. Statistical power is a function of Type II error (or β) and is represented by $1 - \beta$. In the context of model fit, Type II errors occur when a measure of fit fails to detect the misfit and therefore, fails to reject the misspecified model. As such, statistical power of fit indices represents its probability of rejecting a poor fitting model when it is truly misspecified. Toward this end, hit rates which are a function of true positives (TP) and false negatives (FN) were determined as: $\frac{TP}{TP+FN}$. In sum, hit rates correspond to whether the model was correctly rejected when consulting specific criteria or cut-offs.

For the purposes of this study, the following cut-offs were used: 0.012 for RMSEA, 0.979 for TLI, and 0.987 for Mc. In terms of structural fit indices, the cut-offs were: 0.017 for RMSEA-P, 0.994 for C9, 0.006 for C10. It is clear that the cut-offs utilized in this study depart from those that are traditionally used [see Schermelleh-Engel et al. (2003)]. Specifically, these cut-offs were derived in a previous study (Rifenbark, 2019) via a simulation in which the magnitude of factor loadings, the number of indicators per factor, and group sample sizes were systematically varied. The cut-offs were selected such that they correspond to the 95th percentile. In terms of $\Delta\chi^2$ and χ^2 critical values that correspond to an alpha of 0.05 were utilized.

Afterwards, univariate ANOVAs were estimated to determine the impact of design factors on the performance of the various measures of fit. Specifically, between-subject factors were: model size, magnitude of factor loadings, reference group sample size, and severity of misspecifications; additionally, all possible interactions were entered into the model. From these models partial η^2 was consulted and 0.01, 0.06, and 0.14 were considered to be small, medium, and large effect sizes, respectively (Cohen, 1988).

4.2 Results

A convergence rate of 100% was achieved across all 108 simulation conditions. A subset of the study results are presented here and interested readers can view all simulation results on the author's [OSF account](#). Further, due to the similar performance between the non-centrality [$\chi^2 - df$] and ratio based [$\frac{\chi^2}{df}$] approaches for constructing the C9 and C10 SFIs, only the non-centrality approach is reported hereinafter.

Mean Misspecification After applying the empirically derived cut-off values from Rifenbark (2019) across simulation conditions, mean hit rates were estimated over all 54,000 replications. With respect to structural fit measures, overall power was estimated to be: 0.77 (SD = 0.42) for both the C9 and C10 SFIs, 0.82 (SD = 0.39) for RMSEA-P, and 0.83 (SD = 0.38) for $\Delta\chi^2$. In terms of global measures of fit, overall power rates were found to be lower, specifically: 0.47 (SD = 0.50) for TLI, 0.59 (SD = 0.49) for Mc, 0.71 (SD = 0.46) for RMSEA, and 0.68 (SD = 0.47) for χ^2 . Next, an analysis of variance was estimated for each of the measures of fit using the hit rates as the dependent variable with the following factors: model size, magnitude of factor loading, reference group sample size, and severity of the misspecification, as well as all possible interactions. Afterwards, partial η^2 was selected as an effect size to determine the impact of these factors on power for detecting structural misspecification.

In terms of structural measures of fit, a small-to-medium effect was found for both C9 and C10 for the interaction between magnitude of factor loadings and severity (partial $\eta^2 = 0.04$) and a medium effect was observed for both RMSEA-P and $\Delta\chi^2$ (partial $\eta^2 = 0.07$). With respect to global fit measures, a small effect was observed for TLI for the interaction between magnitude of factor loadings, reference group sample size, and severity (partial $\eta^2 = 0.03$); meanwhile a large effect was observed for the interaction between magnitude of factor loadings and severity (partial $\eta^2 = 0.14$). Additionally a large effect was observed for Mc for the interaction between magnitude of factor loadings and severity (partial $\eta^2 = 0.20$); whereas, a small effect was observed for RMSEA (partial $\eta^2 = 0.03$) and a medium effect was observed for χ^2 (partial $\eta^2 = 0.06$).

Covariance Misspecification Upon converting fit measure estimates into hit rates based on cut-off derived by Rifenbark (2019), mean power estimates were computed across all replications. In terms of structural measures of fit, overall power was 0.74 (SD = 0.44) for both the C9 and C10, 0.76 (SD = 0.43) for RMSEA-P, and 0.77 (SD = 0.42) for $\Delta\chi^2$. With respect to global fit indices, overall power was estimated to be 0.25 (SD = 0.44) for TLI, 0.42 (SD = 0.49) for Mc, 0.57 (SD = 0.50) for RMSEA, and 0.54 (SD = 0.50) for χ^2 . In a similar fashion, analysis of variance was conducted for each measure and its hit rates to determine the impact of study factors on power rates.

In terms of structural measures of fit, a small-to-medium effect was observed for the interaction between magnitude of factor loadings and severity for the C9 and C10 SFIs (partial $\eta^2 = 0.04$); whereas, a large effect was observed for both the RMSEA-P and $\Delta\chi^2$ (partial $\eta^2 = 0.14$). In terms of global fit measures, a small effect was observed for the interaction between magnitude of factor loadings, reference group sample size, and severity for RMSEA (partial $\eta^2 = 0.01$). For the TLI, two interaction effects were observed, first, a medium effect for the interaction between magnitude of factor loadings and severity (partial $\eta^2 = 0.06$) and

a small effect for the interaction between reference group sample size and severity (partial $\eta^2 = 0.01$). For Mc, a large effect was observed for the interaction between magnitude of factor loadings and severity (partial $\eta^2 = 0.26$); whereas a medium-to-large effect was observed for χ^2 (partial $\eta^2 = 0.10$).

Visualizing the Impact of Construct Reliability Depending on the measure of fit, partial η^2 estimates differed for the main effects and the interaction effects. Therefore, plots were created to graphically represent the effect study design factors had on power rates and to what extent power rates differed depending on whether the mean or covariance structure was misspecified. To be succinct, power plots are presented for C9 (see Fig. 2), RMSEA-P (see Fig. 3), and TLI (see Fig. 4).

5 Discussion

This study was conducted to understand the impact facets of construct reliability have on statistical power for measures of fit to detect a misspecified structural model. This was done in the context of multiple group SEM where systematic misspecifications were placed in either the mean or covariance structure with varying levels of severity by way of true population differences on targeted structural parameters between groups. Therefore, when all structural parameters were constrained to be the same across groups, structural misfit was introduced. Across all simulation conditions, common GFIs on average failed to detect the misspecification of the structural model regardless of whether the covariance or mean structure was misspecified when consulting Hu and Bentler (1998, 1999) cut-offs, this confirms previous research (Heene et al., 2011; Lance et al., 2016; Hancock & Mueller, 2011). For this information, please see the supplemental material housed on OSF.

When viewing the power plots for the selected measures of fit, a pattern emerges. Primarily, as the magnitude of factor loadings increase given a misspecified structural model, power to detect the misspecification also increases. However, there are two exceptions: C9 given a small mean misspecification and TLI regardless of whether it was the mean or covariance structure that was misspecified. Interestingly, the impact of the magnitude of the factor loadings on power rates for TLI was the inverse, namely as factor loadings decrease the TLI seems to possess more statistical power to detect the structural misspecification. It is believed that this behavior has to do with the TLI's reliance on a baseline (or null model). Another pattern that emerged has to do with the difference in power rates when the covariance structure was misspecified compared to when the mean structure was misspecified. Specifically, it appears that all measures of fit possessed more power to detect a misspecified mean structure rather than a misspecified covariance structure; however, it is important to recall that different Cohen's d was utilized—based on

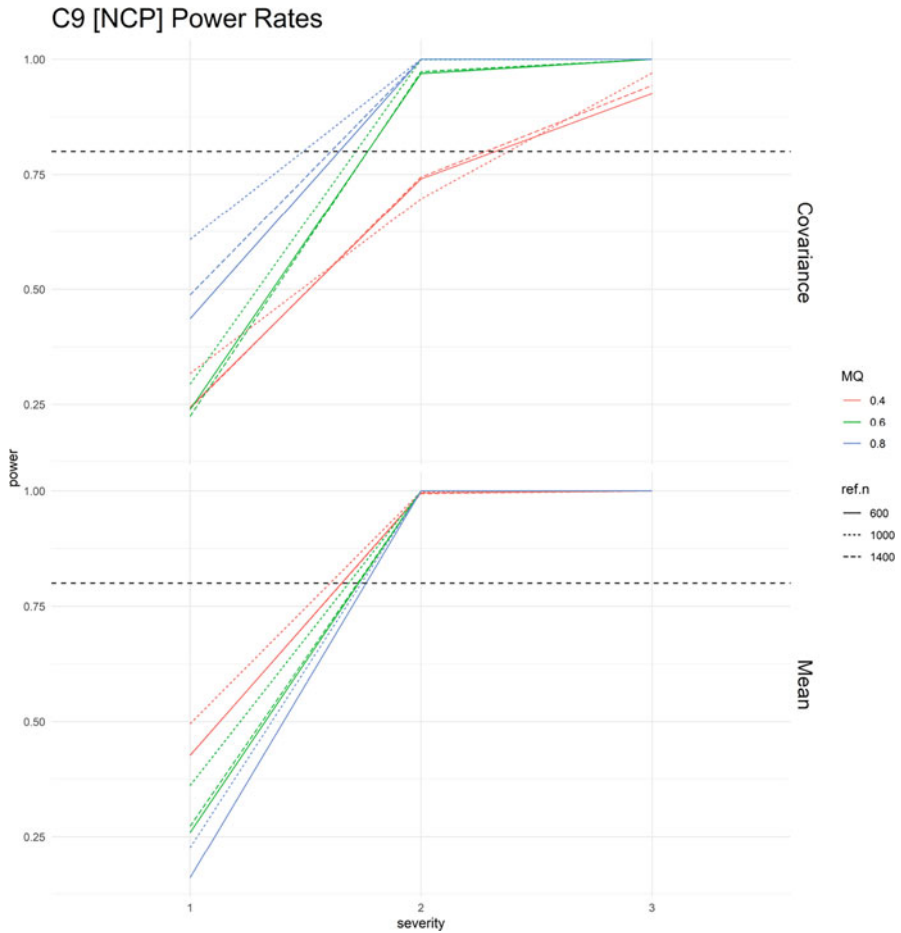


Fig. 2 Power for C9 to detect a mean or covariance misspecification based on severity of misspecification, measurement quality, and reference group sample size. *Note.* MQ = measurement quality (i.e., magnitude of factor loadings); ref.n = reference group sample size. Top corresponds to a misspecified mean structure and the plot on the bottom corresponds to when the covariance structure was misspecified. Line color: Red ($\Lambda = 0.4$), green ($\Lambda = 0.6$), and blue ($\Lambda = 0.8$). Line Type: Solid line (ref.n = 600), dotted (ref.n = 1000), and dashed (ref.n = 1400)

literature—for small, medium, and large misspecification depending on where the misfit was introduced therefore, little weight should be placed on this behavior. Finally, it was found that the magnitude of the factor loadings seemed to have a larger impact on power rates than model size and this was more clear when evaluating the performance of measures of fit given a misspecified covariance structure.

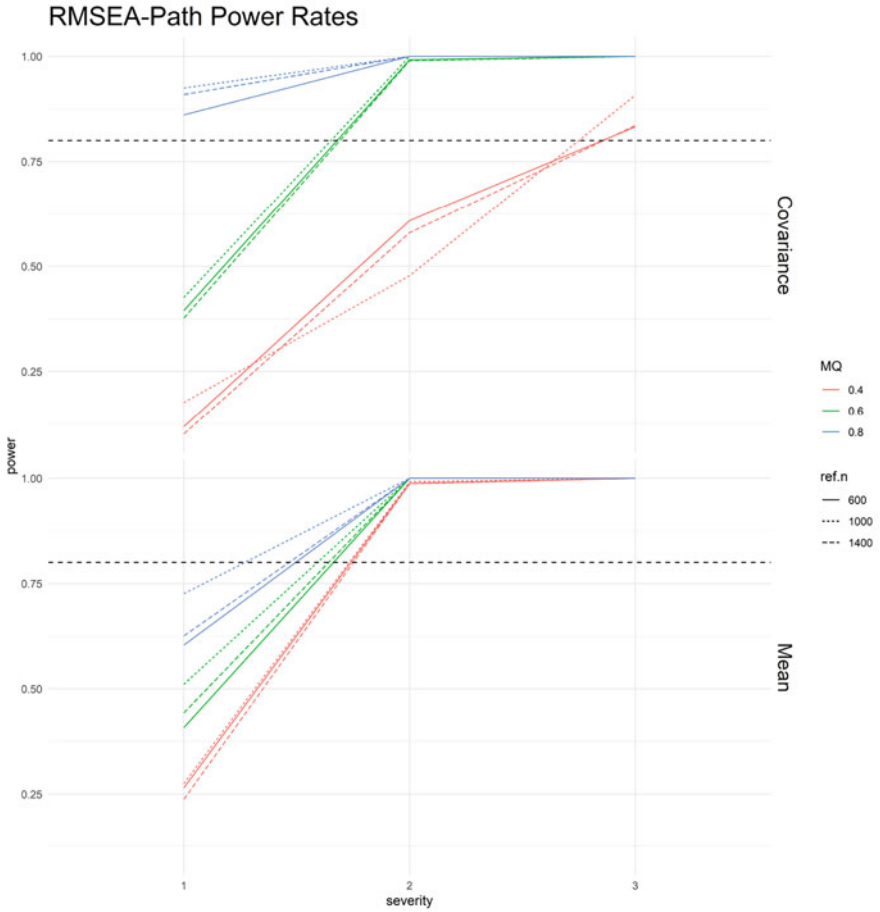


Fig. 3 Power for RMSEA-P to detect a mean or covariance misspecification based on severity of misspecification, measurement quality, and reference group sample size. *Note.* MQ = measurement quality (i.e., magnitude of factor loadings); ref.n = reference group sample size. Top corresponds to a misspecified mean structure and the plot on the bottom corresponds to when the covariance structure was misspecified. Line color: Red ($\Lambda = 0.4$), green ($\Lambda = 0.6$), and blue ($\Lambda = 0.8$). Line Type: Solid line (ref.n = 600), dotted (ref.n = 1000), and dashed (ref.n = 1400)

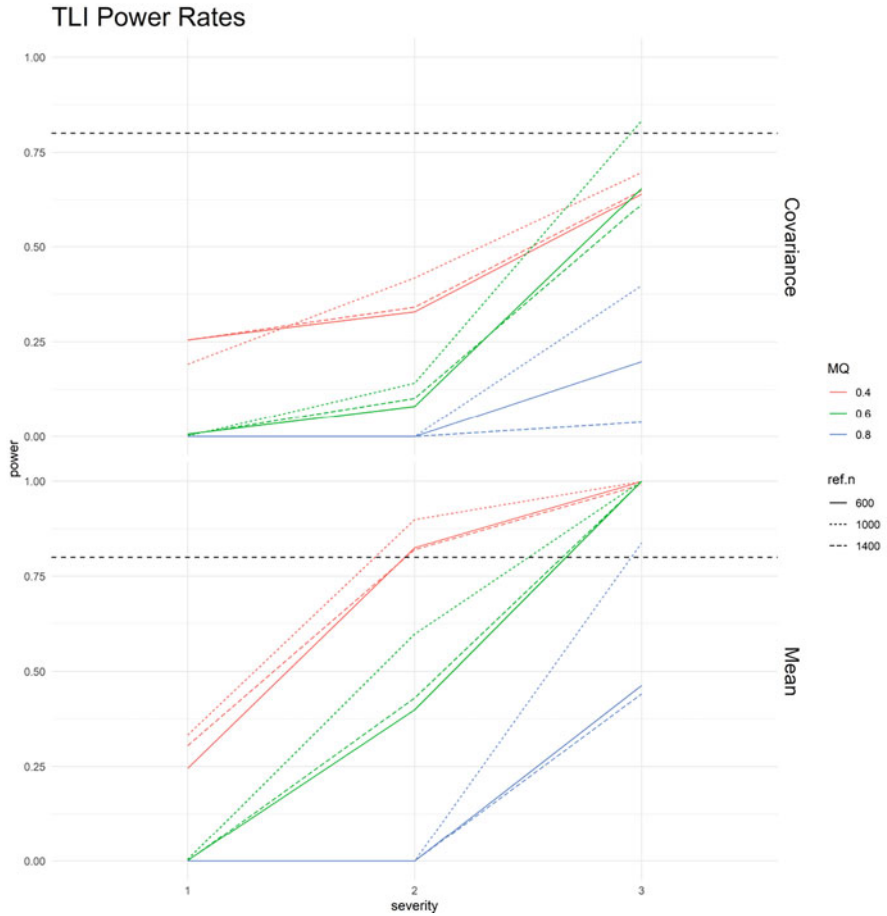


Fig. 4 Power for TLI to detect a mean or covariance misspecification based on severity of misspecification, measurement quality, and reference group sample size. *Note.* MQ = measurement quality (i.e., magnitude of factor loadings); ref.n = reference group sample size. Top corresponds to a misspecified mean structure and the plot on the bottom corresponds to when the covariance structure was misspecified. Line color: Red ($\Lambda = 0.4$), green ($\Lambda = 0.6$), and blue ($\Lambda = 0.8$). Line Type: Solid line (ref.n = 600), dotted (ref.n = 1000), and dashed (ref.n = 1400)

References

- Browne, M. W., & Arminger, G. (1995). Specification and estimation of mean-and covariance-structure models. In *Handbook of statistical modeling for the social and behavioral sciences* (pp. 185–249). New York: Springer.
- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Mahwah: Lawrence Erlbaum Associates, Publishers.
- Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling: A Multidisciplinary Journal*, 2(2), 119–143.
- Epskamp, S., & with contributions from Simon Stuber (2017). *semPlot: Path Diagrams and Visual Analysis of Various SEM Packages' Output*. R package version 1.1.
- Fan, X., & Sivo, S. A. (2009). Using δ goodness-of-fit indexes in assessing mean structure invariance. *Structural Equation Modeling*, 16(1), 54–69.
- Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39–50.
- Gagne, P., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*, 41(1), 65–83.
- Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement*, 71(2), 306–324.
- Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, 16(3), 319.
- Hu, L.-t., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3(4), 424.
- Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55.
- James, L., Mulaik, S., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data*. Beverly Hills, CA: Sage.
- Kang, Y., McNeish, D. M., & Hancock, G. R. (2016). The role of measurement quality on practical guidelines for assessing measurement and structural invariance. *Educational and Psychological Measurement*, 76(4), 533–561.
- Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of rmsea in models with small degrees of freedom. *Sociological Methods & Research*, 44(3), 486–507.
- Lance, C. E., Beck, S. S., Fan, Y., & Carter, N. T. (2016). A taxonomy of path-related goodness-of-fit indices and recommended criterion values. *Psychological Methods*, 21(3), 388.
- MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, 100(1), 107.
- McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, 7(1), 64.
- Pornprasertmanit, S., Miller, P., & Schoemann, A. (2016). *simsem: SIMulated Structural Equation Modeling*. R package version 0.5-13.
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Riftenbark, G. G. (2019). *Misfit at the Intersection of Measurement Quality and Model Size: A Monte Carlo Examination of Methods for Detecting Structural Model Misspecification*. PhD thesis, University of Connecticut.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36.

- Schermelleh-Engel, K., Moosbrugger, H., Müller, H., et al. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2), 23–74.
- Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, 27(2), 229–239.
- Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, 8(1), 16.

Index

A

Aberrant examinee behavior, 90, 91
Aberrant responses, 84
Absolute deviations of observed and predicted corrections, 18–20, 23
Attribute-balancing, 302, 305, 307, 309
Autoregressive model, 185–201

B

Bayesian statistics, 112
Bullying, 96–100, 102–107

C

Categorical data, 185–201
Chi-squared statistic, 18–20, 22, 23, 246
Classical test theory (CTT), 11, 52, 53, 62, 161–162, 220–223, 229, 233
Coefficient α , 52, 55–62
Cognitive diagnosis models, 17–24
Cognitive diagnostic computerized adaptive testing (CD-CAT), 299–301, 309
Cognitive diagnostic models (CDMs), v, 22, 24, 84, 165, 217–224, 300–301, 309
College admissions test, 41, 42, 47
Conditional likelihood, 112, 114, 116, 118, 119, 186, 190, 191, 194, 195
Construct reliability, 313–326
Continuation ratio model, 96–107

D

Deviance information criteria (DIC), 113–119
Differential item functioning (DIF), 68, 107, 269, 278–285

Digital-first assessment, 265–275
Discretization, 227–234
Distractor quality, 91

E

Educational topic models, 123, 288
Exact likelihood, 186, 189, 194, 195

F

Factor analysis, 52–62, 221, 228
Feedback, 218, 237–247
Forensic science, 72, 80, 81
Formative assessment, 238, 247

G

Gibbs sampling, 28, 30–33, 35, 36, 207–209, 258
Greatest lower bound (GLB), 52, 53, 55–62

H

High-stakes assessment, 85, 266, 274

I

Individualized learning, 121–138
Item characteristic curve (ICC), 12, 13, 171–183
Item difficulty modeling (IDM), 171, 172
Item exposure control, 307
Item response modeling, 12, 96, 171

Item response theory (IRT), v, 40, 52, 53, 62, 67, 68, 70, 80, 84, 96–99, 106, 161–164, 172, 173, 220–223, 269, 278, 280

L

Latent Dirichlet allocation (LDA), 27–36, 122, 123, 203–214

Leave-one-out cross validation (LOO-CV), 113–119

Linear normal one-factor (LNF) model, 227–234

Lord's chi-square, 278, 280–282, 285

M

Machine learning, 1, 2, 7, 164, 171, 173

Mantel-Haenszel (MH) statistic, 18–20, 22, 23

Marginal likelihood, 30, 112–116, 118, 119

Mathematics education, 289

Maximum likelihood estimation (MLE), 188–192, 195, 258, 278–285, 301, 303, 304, 310, 311, 320

Measurement efficiency, 304

Metadata, 288, 289, 297

Method of percentiles (MOP), 142, 149

Missing data, 106, 252, 277–285

Model recovery, 204, 205, 208, 211, 213, 258

Multiple-choice test, 66, 83

Multiple group, 315, 323

Multiple imputation (MI), 278–285, 304

N

Natural language processing (NLP), 90, 151–157, 164, 171

Neural networks (NNs), 153, 154, 172–177, 179, 180, 183

O

Optimal scores, 40–48

P

Pólya frequency functions, 229

Polytomous items, 55, 95–107

Pooling time series, 186, 195

Power method, 142

Predictive validity, 40, 41, 46, 47, 229

Q

Q-matrix, 18, 20–22, 113, 217–224, 300, 302, 304

Quality assurance, 14, 265–275

R

Recommendation system, 121–138

Reliability, 1–14, 51–62, 66, 72, 106, 152, 165, 221–223, 269, 313–326

Response time, v, 4, 5, 66, 68–70, 85–90, 269, 299–311

Retrofit, 217–224

Round-robin designs, 250

S

Semantic similarity, 177–179, 183

Simulation study, 20–23, 28, 31, 36, 52, 54–62, 113, 115–119, 135, 186, 191–192, 204, 205, 208, 209, 211–213, 233, 261, 297, 307, 308, 315, 318–323

Social networks, 256, 257

Social relations model (SRM), 249–262

Statistical power, 280, 281, 285, 321, 323

Structural equation modeling (SEM), 315, 316

Structural misspecification, 314, 315, 319, 320, 322, 323

Structural topic model (STM), 288–294, 296, 297

Sum scores, 5, 10, 11, 40–48, 106, 107, 135, 227–234

Supervised Dirichlet allocation, 288, 289

T

Testlet effects, 17–24

Topic model, 28, 29, 122–126, 129, 135, 204, 288, 289

Topic modeling, 121–138, 210, 287–297

Totally positive densities, 229

Tukey distribution, 141–149

U

Unfolding model, 238, 240, 245, 247

V

Variational expectation maximization (VEM), 28, 30–33, 35

W

Widely-applicable information criterion (WAIC), 70, 71, 77, 113–119

Writing, 220, 238, 241, 243–245, 247