Springer Proceedings in Mathematics & Statistics

L. Andries van der Ark
Daniel M. Bolt
Wen-Chung Wang
Jeffrey A. Douglas
Marie Wiberg *Editors*

Quantitative Psychology Research

The 80th Annual Meeting of the Psychometric Society, Beijing, 2015

Springer

Springer Proceedings in Mathematics & Statistics

Volume 167

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

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of select contributions from workshops and conferences in all areas of current research in mathematics and statistics, including OR 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.

L. Andries van der Ark • Daniel M. Bolt
Wen-Chung Wang • Jeffrey A. Douglas
Marie Wiberg
Editors

Quantitative Psychology Research

The 80th Annual Meeting of the Psychometric Society, Beijing, 2015

Springer

Editors
L. Andries van der Ark
University of Amsterdam
Amsterdam, The Netherlands

Daniel M. Bolt
University of Wisconsin
Madison, Wisconsin, USA

Wen-Chung Wang
Education University of Hong Kong
Hong Kong, China

Jeffrey A. Douglas
University of Illinois
Champaign, Illinois, USA

Marie Wiberg
Umeå University
Umeå, Sweden

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-38757-4 ISBN 978-3-319-38759-8 (eBook)
DOI 10.1007/978-3-319-38759-8

Library of Congress Control Number: 2016944495

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

This volume represents presentations given at the 80th annual meeting of the Psychometric Society, organized by the Beijing Normal University, during July 12–16, 2015. The meeting attracted 511 participants from 21 countries, with 254 papers being presented, along with 119 poster presentations, three pre-conference workshops, four keynote presentations, eight invited presentations, and six invited and five contributed symposia. This meeting was the first ever held in China, the birthplace of standardized testing, as was highlighted in the keynote address "the history in standardized testing" by Dr. Houcan Zhang. We thank the local organizers Tao Xin and Hongyun Liu and their staff and students for hosting this very successful conference.

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

We asked authors to use their presentation at the meeting as the basis of their chapters, possibly extended with new ideas or additional information. The result is a selection of 29 state-of-the-art chapters addressing a diverse set of topics, including item response theory, factor analysis, structural equation modelling, time series analysis, mediation analysis, cognitive diagnostic models, and multi-level models.

Amsterdam, The Netherlands — L. Andries van der Ark
Madison, WI — Daniel M. Bolt
Hong Kong, China — Wen-ChungWang
Urbana-Champaign, IL — Jeffrey A. Douglas
Umeå, Sweden — Marie Wiberg

Contents

Continuation Ratio Model in Item Response Theory and Selection of Models for Polytomous Items

Seock-Ho Kim

Abstract In the continuation ratio model continuation ratio logits are used to model the probabilities of obtaining ordered categories in polytomously scored items. The continuation ratio model is an alternative to other models for ordered category items such as the graded response model, the generalized partial credit model, and the partial credit model. The theoretical development of the model, descriptions of special cases, maximum likelihood estimation of the item and ability parameters are presented. An illustration and comparisons of the models for ordered category items are presented using empirical data.

Keywords Bayesian estimation • Continuation ratio model • Item response theory • Maximum likelihood estimation • Multicategory logit model • Polytomous model

1 Introduction

When a free response item is scored in a dichotomous fashion, a single decision is performed in a sense that no further decisions will be made beyond the current decision to be taken. When a free response item is rated in a polytomous fashion, either a single decision is performed or multiple decisions in which dependent decisions are made in tandem are required.

Borrowing terms from the game theory (Luce & Raiffa, 1957), a particular alternative chosen by a rater at a given decision point is called a "choice," and the totality of choices available to a rater at the decision point constitutes a "move." A sequence of choices, one following another until the rating or scoring of an item is complete, can be called a "play." The play or the rating process for a given item can be depicted with a connected graph, called a decision tree, consists of a collection of nodes and branches between pairs of nodes. A decision tree with three decision points and four choices is presented in Fig. 1. The decision tree reflects

S.-H. Kim (✉)
Department of Educational Psychology, The University of Georgia, 325 Aderhold Hall, Athens, GA 30602-7143, USA
e-mail: shkim@uga.edu

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_1

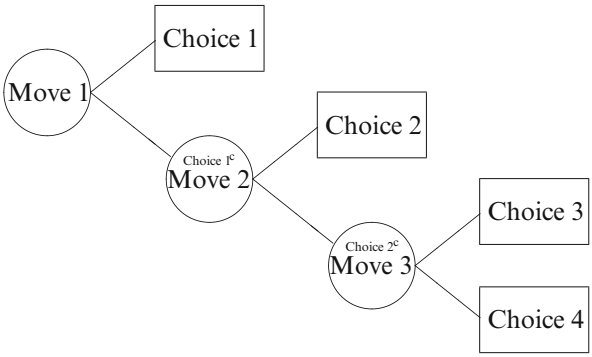


Fig. 1 A decision tree with three decision points and four choices

the sequential nature of scoring. Each decision point is denoted as a circle and the chance events with respective but dependent probabilities are denoted as squares in Fig. 1. The superscript c of the choice number indicates the complement of the event.

The decision tree in Fig. 1 involves in a set of dependent events. The model for the ordered choices ought to reflect the joint probabilities and must take into account the conditional probabilities that characterize the dependence. The model for ordered category items to be described is called a continuation ratio model. Such a model that employs continuation-ratio logits with a manifested or directly-observed explanatory variable was originally developed to handle a multicategory response variable in logit models (Cox 1972). In the item response theory field, Mellenbergh (1995) presented conceptual notes on models for discrete polytomous item responses and indicated that the continuation ratio model could be considered as a special case of the Bock's (1972) model (cf. Tutz 1990; Hemker, van der Ark, & Sijtsma, 2001). The general discussion of the various item response theory models for polytomously scored items can be found in Hambleton, van der Linden, and Wells (2010).

2 The Continuation Ratio Model and Parameter Estimation

Let Y_{ij} be a random variable that designates the rating or scored item response of individual i to item j. The continuation ratio model assumes that the manifestation of Y_{ij} or the probability of Y_{ij} to be a specific value depends on a person's latent ability θ_i and a vector-valued item characteristic ξ_j [i.e., a_{jk}'s and b_{jk}'s; see the definitions following Eq. (1)]. The probability that $y_{ij} = k$ given ability θ_i and item parameter ξ_j, Prob $\left(y_{ij} = k|\theta_i, \xi_j\right)$, is

$$P_{jk}(\theta_i) = \begin{cases} \dfrac{\exp\left[-a_{jk}(\theta_i - b_{jk})\right]}{\prod_{h=1}^{k}\left\{1+\exp\left[-a_{jh}(\theta_i - b_{jh})\right]\right\}} & \text{for } k = 1, \ldots, K_j - 1 \\ \dfrac{1}{\prod_{h=1}^{K_j-1}\left\{1+\exp\left[-a_{jh}(\theta_i - b_{jh})\right]\right\}} & \text{for } k = K_j, \end{cases} \tag{1}$$

where a_{jk} is the slope parameter and b_{jk} is the threshold parameter. The number of item parameters for item j is $2(K_j - 1)$. When an item has two rating categories, that is, $K_j = 2$, the continuation ratio model becomes the two-parameter logistic model.

Under the assumption of conditional independence, the probability of a response vector $y_i = (y_{i1}, \ldots, y_{iJ})$, is given as $\text{Prob}\,(y_i|\theta_i, \xi) = p(y_i|\theta_i, \xi_1, \ldots, \xi_J) = \prod_{j=1}^{J} P_{jk}(\theta_i)$ and the joint probability of the response vectors of a sample of I subjects is given as $\text{Prob}\,(y|\theta, \xi) = p(y_1, \ldots, y_I|\theta_1, \ldots, \theta_I, \xi) = \prod_{i=1}^{I}\prod_{j=1}^{J} P_{jk}(\theta_i)$. When the joint probability is considered as a function of unknown parameters ξ and θ, we call it the likelihood L. Inference of the values of unknown parameters from observed data can be accomplished by maximizing the likelihood or its modifications with respect to the unknown parameters.

Several estimation procedures are available to obtain parameter estimates in the continuation ratio model. Kim (2002) presented detailed estimation procedures including the marginal estimation of item parameters (Bock & Aitkin, 1981). Kim (2002) also presented model fit statistics, estimation of the latent criterion variable θ_i (i.e., methods of maximum likelihood, maximum a posteriori, and expected a posteriori), and information functions for the continuation ratio model.

It can be noted that the continuation ratio model treats a polytomously scored item as a set of dichotomously scored items (Kim 2013). For example, an item with four categories or choices can be converted into three dichotomously scored items with some dependency among the converted dichotomous items. It is possible, consequently, to obtain the parameter estimates under the continuation ratio model using computer programs that implemented the marginal maximum likelihood estimation of item parameters under the usual two-parameter logistic model and an ability estimation method. Kim (2013) presented means to obtain parameter estimates using several popular item response theory computer programs utilizing missing or not-presented options.

Note that other parameter estimation methods (e.g., Bayesian estimation, Markov chain Monte Carlo, Gibbs sampling; see Baker & Kim, 2004) implemented in item response theory computer programs can also be applied to obtain both item and ability parameter estimates under the continuation ratio model. Because of the relationship between the two-parameter logistic model and the continuation ratio model, priors of item parameters used in Bayesian estimation can be employed with minor changes (e.g., Swaminathan & Gifford 1985).

Although the continuation ratio model for the polytomous items with ordered categories has been available for some time, applications of the model to analyze

polytomous data are not widely available. An illustration is presented next using empirical data with the Fortran implementation of the continuation ratio model and the computer program MULTILOG (Thjssen, Chen, & Bock, 2002). Subsequently, comparisons of the estimation results from several models for ordered category items are presented using MULTILOG.

3 An Illustration

The data from an experimental form of a French writing assessment were analyzed. The experimental form was a performance assessment rating instrument that consists of three polytomously scored items with four ordered rating categories. The participants were 120 college students who had complete data for the three item responses. Although there might be 64 different response patterns, 31 distinctive patterns were actually observed (see Table 2 for the response patterns and the number of examinees in each pattern).

The marginal maximum likelihood estimation of item parameters was carried out on the three French items from the experimental form using the Fortran computer program modified from the code written for Kim (2002). Ten quadrature fractile points were used for ability integration during calculations. After several cycles of the expected and maximization iterations, the item parameter estimates were stable to four significant figures. Goodness of fit for the model was assessed, and the resulting chi-square value of the -2 log likelihood was 38.81 with the degrees of freedom of 12 (i.e., the number of response patterns minus the number of parameters estimated minus one; see Bock & Aitkin, 1981). Although the solution showed reasonably good fit, the chi-square was relatively large (i.e., $p < .01$) due to the sparseness of data from the small frequencies of the 31 observed response patterns. Ability parameters were estimated with a method of expected a posteriori (Bock & Mislevy, 1982) using the Fortran program written for Kim (2002).

Item and ability parameter estimates of the continuation ratio model from MULTILOG were also obtained. The input files for the MULTILOG run are shown in the Appendix (i.e., FRENF.MLG and the data file without a name, e.g., FRENF.DAT). The exact interpretation of the keywords and command lines can be found in the manual of the computer program MULTILOG (see Thissen et al. 2002; du Toit 2003).

Item parameter estimates and standard errors of the continuation ratio model from the Fortran implementation of the marginal maximum likelihood estimation as well as those from MULTILOG are presented in Table 1. Because the source code of the proprietary program is not in general available, the estimation result from the Fortran implementation based on open source (i.e., the Fortran source code is available from the author) was used here as a reference purpose. All of the item parameter estimates for a given item between two computer programs are very similar. It should be noted that by changing the default settings of the program, it may be possible to obtain exactly the same estimation results.

Table 1 The continuation ratio model item parameter estimates and standard errors (s.e.) from the Fortran program and MULTILOG

Program	Item	Item parameter estimate					
		a_{j1} (s.e.)	b_{j1} (s.e.)	a_{j2} (s.e.)	b_{j2} (s.e.)	a_{j3} (s.e.)	b_{j3} (s.e.)
Fortran	1	2.22 (0.65)	−1.36 (0.25)	2.60 (1.01)	−0.09 (0.26)	3.89 (1.77)	1.34 (0.17)
	2	2.59 (1.32)	−1.85 (0.37)	2.72 (1.25)	−0.31 (0.14)	3.61 (1.11)	1.12 (0.18)
	3	2.14 (0.43)	−1.55 (0.25)	1.79 (0.48)	−0.34 (0.19)	3.91 (1.33)	0.92 (0.17)
MULTILOG	1	2.17 (0.61)	−1.38 (0.27)	2.43 (0.67)	−0.11 (0.15)	3.68 (1.24)	1.32 (0.18)
	2	2.68 (1.06)	−1.84 (0.32)	2.96 (0.79)	−0.31 (0.13)	3.82 (1.28)	1.11 (0.15)
	3	2.17 (0.52)	−1.55 (0.29)	1.71 (0.47)	−0.37 (0.22)	4.49 (1.54)	0.93 (0.13)

Plots of the category response functions of the three items under the continuation ratio model were obtained and presented in Fig. 2. For each of the items, the monotonic decreasing curve corresponds to the lowest category; the middle two curves correspond to the two middle categories; the monotonic increasing curve corresponds to the highest category. These indicate in each item that the examinees of indefinitely low ability will be assigned the lowest category and, conversely, that examinees of indefinitely high ability will be assigned the highest category. Considering the size of standard errors, these differences may be trivial. In sum, all category response functions from the programs are nearly the same, reflecting the similarity in the item parameter estimates.

Ability estimates from the method of expected a posteriori assuming that item parameter estimates under the continuation ratio model from the Fortran implementation to be true values were obtained and reported in Table 2. Ability estimates were also obtained from MULTILOG. A standard normal prior was used in ability estimation. Due to the similarity of the item parameter estimates, the ability estimates are very similar. One peculiar ability estimate was obtained for the response pattern of 443. The ability estimate was less than those obtained from the response patterns of 441 and 442. A procedure or constraint to prevent to yield illogical ability estimates may be applied in practice.

4 Comparisons of Polytomous Models

The same data from the experimental form of the French writing assessment were analyzed to compare models for ordered category items. Category response functions of the items under the graded response model (Samejima 1969), the generalized partial credit model (Muraki 1992), and the partial credit model (Masters 1982) were obtained using MULTILOG. Example input files for various polytomous models can be found in du Toit (2003).

Item parameter estimates under the graded response model, the generalized partial credit model, and the partial credit model are reported in Table 3. It should

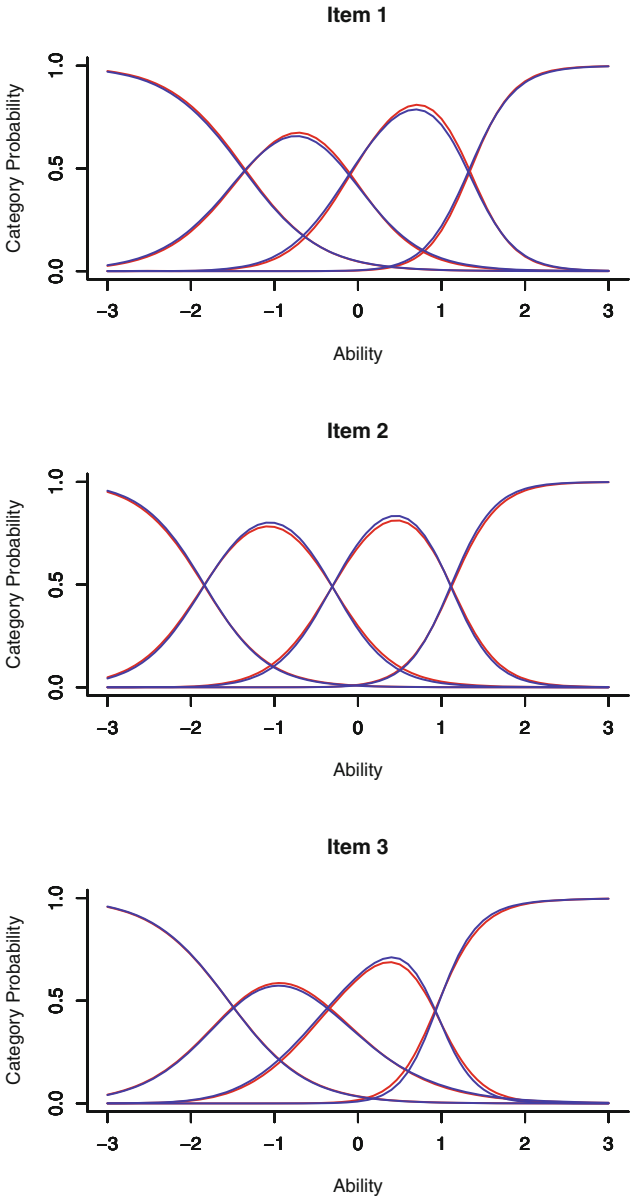


Fig. 2 Category response functions for items 1–3 under the continuation ratio model from the Fortran program (*red*) and MULTILOG (*blue*)

be noted that the actual, unconstrained parameters estimated in the generalized partial credit model and the partial credit model from MULTILOG are those under the nominal response model. The output from MULTILOG contained both

Table 2 Expected a posteriori (EAP) ability estimates and the posterior standard deviations (p.s.d.) from the Fortran program and MULTILOG

		Program				Program	
		Fortran	MULTILOG			Fortran	MULTILOG
Pattern	n	EAP (p.s.d.)	EAP (p.s.d.)	Pattern	n	EAP (p.s.d.)	EAP (p.s.d.)
111	4	−2.10 (0.57)	−2.10 (0.53)	233	9	0.01 (0.49)	0.02 (0.40)
112	1	−1.56 (0.44)	−1.61 (0.47)	323	6	−0.00 (0.49)	−0.05 (0.41)
121	4	−1.47 (0.41)	−1.47 (0.47)	332	4	0.23 (0.45)	0.18 (0.42)
211	1	−1.53 (0.44)	−1.58 (0.48)	234	1	0.53 (0.28)	0.69 (0.36)
122	4	−1.14 (0.49)	−1.09 (0.45)	243	1	0.51 (0.27)	0.60 (0.37)
212	1	−1.19 (0.48)	−1.17 (0.46)	333	24	0.42 (0.27)	0.37 (0.37)
221	4	−1.08 (0.50)	−1.07 (0.45)	342	2	0.73 (0.45)	0.83 (0.39)
123	2	−0.70 (0.47)	−0.75 (0.45)	423	1	0.55 (0.31)	0.54 (0.39)
132	1	−0.51 (0.43)	−0.48 (0.46)	441	1	1.41 (0.34)	1.31 (0.41)
222	10	−0.69 (0.43)	−0.76 (0.42)	334	5	0.74 (0.43)	0.89 (0.31)
312	1	−0.55 (0.46)	−0.63 (0.49)	343	1	0.69 (0.40)	0.82 (0.32)
223	5	−0.46 (0.33)	−0.46 (0.41)	442	1	1.46 (0.32)	1.40 (0.42)
232	8	−0.33 (0.40)	−0.24 (0.42)	344	4	1.40 (0.27)	1.25 (0.33)
322	3	−0.34 (0.39)	−0.33 (0.42)	434	2	1.42 (0.25)	1.24 (0.33)
134	1	0.52 (0.33)	0.66 (0.38)	443	1	1.40 (0.27)	1.17 (0.32)
Continued to the right-hand-side columns				444	7	1.74 (0.48)	1.81 (0.48)

unconstrained item parameter estimates as well as those transformed estimates with Bock's (1972) contrasts. The estimates reported under the generalized partial credit model and the partial credit model in Table 3 are the ones actually estimated by MULTILOG (see du Toit 2003 pp. 570–595).

Plots of category response functions obtained from the MULTILOG runs for the continuation ratio model, and the three other polytomous item response theory models are presented in Fig. 3. The third and fourth category response functions from the continuation ratio model seem different from those from the other polytomous item response theory models. The category response functions for item 2 from the graded response model and the generalized partial credit model look nearly the same.

The full-information fit statistics from MULTILOG were $G^2(12) = 40.4$ for the continuation ratio model, $G^2(22) = 45.5$ for the graded response model, $G^2(22) = 50.6$ for the generalized partial credit model, and $G^2(24) = 51.5$ for the partial credit model. All likelihood-ratio goodness-of-fit statistic values were statistically significant (i.e., $p < .01$) and relatively large due to the sparseness of data.

In addition, the Akaike's (1992) AIC (i.e., an information criterion) was obtained. The AIC values were 791.55 for the continuation ratio model, 784.66 for the graded response theory model, 789.75 for the generalized partial credit model, and 786.67 for the partial credit model (see Kang & Cohen, 2007). The graded response model seems to be the best fitting one for the current data. Thissen, Nelson, Rosa, and

Table 3 Item parameter estimates and standard errors (s.e.) from the graded response (GR) model, the generalized partial credit (GPC) model, and the partial credit (PC) model

GR model	Item	MULTILOG estimate			
		a_j (s.e.)	b_{j1} (s.e.)	b_{j2} (s.e.)	b_{j3} (s.e.)
	1	2.81 (0.45)	−1.26 (0.17)	−0.08 (0.12)	1.46 (0.21)
	2	3.00 (0.51)	−1.75 (0.22)	−0.31 (0.11)	1.19 (0.17)
	3	2.42 (0.35)	−1.47 (0.22)	−0.26 (0.14)	1.15 (0.20)

GPC model	Item	MULTILOG estimate			
		α_j (s.e.)	γ_{j1} (s.e.)	γ_{j2} (s.e.)	γ_{j3} (s.e.)
	1	2.31 (0.42)	−2.84 (0.64)	−0.23 (0.35)	3.42 (0.73)
	2	2.77 (0.54)	−4.81 (0.99)	−0.86 (0.38)	3.27 (0.64)
	3	1.87 (0.31)	−2.66 (0.54)	−0.54 (0.33)	2.22 (0.54)

PC model	Item	MULTILOG estimate			
		α_j (s.e.)	γ_{j1} (s.e.)	γ_{j2} (s.e.)	γ_{j3} (s.e.)
	1	2.27 (0.23)	−2.79 (0.52)	−0.23 (0.35)	3.38 (0.59)
	2	2.27 (0.23)	−4.11 (0.71)	−0.75 (0.36)	2.81 (0.51)
	3	2.27 (0.23)	−3.11 (0.58)	−0.60 (0.34)	2.57 (0.52)

McLeod (2001) reported that the graded response model might fit rating data better than the generalized partial credit model.

Based on the item parameter estimates from the various polytomous item response theory models, the ability parameters were estimated by the method of expected a posteriori using MULTILOG (see Table 4). Ability estimates from the continuation ratio model, the graded response model, the generalized partial credit model, and the partial credit models were very similar. As mentioned in the discussion of Table 2, one peculiar ability estimate was obtained for the response pattern of 443 under the continuation ratio model. Other models for the polytomous items didn't exhibit such an illogical ability estimate.

5 Discussion

The purpose of the present paper was to provide information for the parameter estimation under the continuation ratio model using the Fortran implementation and MULTILOG. An illustration was provided with the performance assessment rating data. Marginal maximum likelihood estimation of item parameters was employed with the method of expected a posteriori for ability estimation. Item parameter estimates from the two programs under the continuation ratio model were very similar, and the ability estimates were also very much alike.

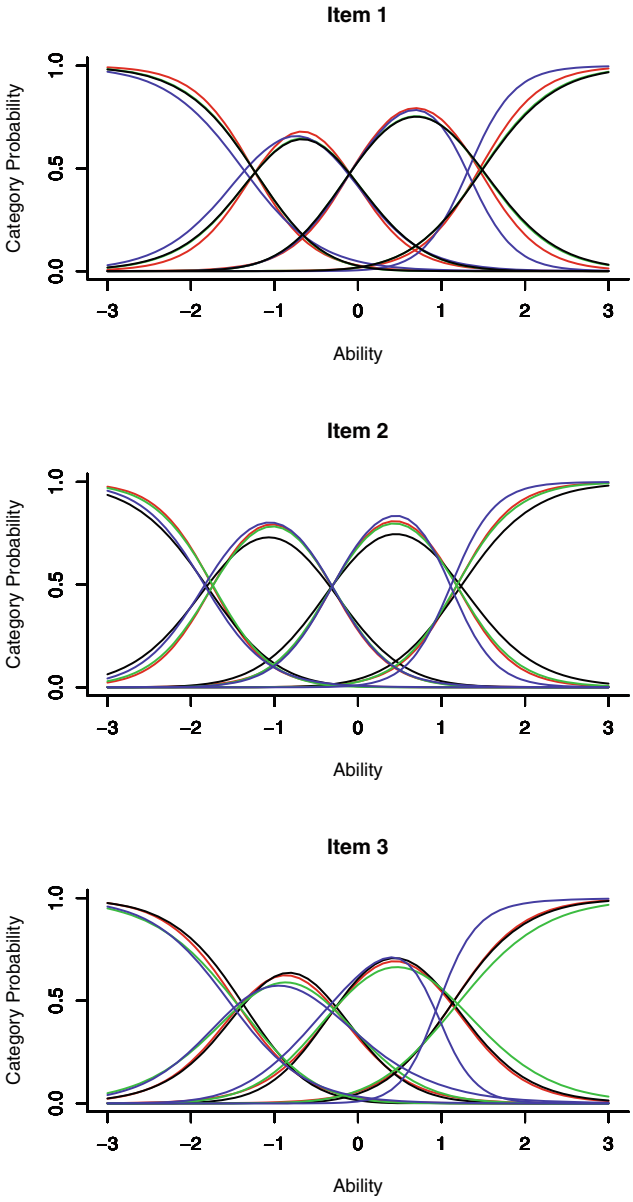


Fig. 3 Category response functions for the continuation ratio model (*blue*), the graded response model (*red*), the generalized partial credit model (*green*), and the partial credit model (*black*)

In addition, the item and ability parameter estimates under the continuation ratio model were compared with those from the graded response model, the generalized partial credit model, and the partial credit model using MULTILOG. Although the

Table 4 Expected a posteriori (EAP) ability estimates and the posterior standard deviations (p.s.d.) under the continuation ratio (CR) model, the graded response (GR) model, the generalized partial credit (GPC) model, and the partial credit (PC) model

		Model			
		CR	GR	GPC	PC
Pattern	*n*	EAP (p.s.d.)	EAP (p.s.d.)	EAP (p.s.d.)	EAP (p.s.d.)
111	4	−2.10 (0.53)	−2.06 (0.51)	−2.05 (0.52)	−2.03 (0.53)
112	1	−1.61 (0.47)	−1.57 (0.43)	−1.61 (0.45)	−1.50 (0.45)
121	4	−1.47 (0.47)	−1.46 (0.42)	−1.44 (0.43)	−1.50 (0.45)
211	1	−1.58 (0.48)	−1.49 (0.43)	−1.52 (0.44)	−1.50 (0.45)
122	4	−1.09 (0.45)	−1.10 (0.40)	−1.10 (0.41)	−1.09 (0.41)
212	1	−1.17 (0.46)	−1.11 (0.41)	−1.18 (0.42)	−1.09 (0.41)
221	4	−1.07 (0.45)	−1.02 (0.41)	−1.03 (0.41)	−1.09 (0.41)
123	2	−0.75 (0.45)	−0.79 (0.44)	−0.79 (0.40)	−0.71 (0.40)
132	1	−0.48 (0.46)	−0.64 (0.43)	−0.64 (0.40)	−0.71 (0.40)
222	10	−0.76 (0.42)	−0.73 (0.48)	−0.72 (0.40)	−0.71 (0.40)
312	1	−0.63 (0.49)	−0.73 (0.48)	−0.79 (0.40)	−0.71 (0.40)
223	5	−0.46 (0.41)	−0.42 (0.39)	−0.42 (0.40)	−0.34 (0.40)
232	8	−0.24 (0.42)	−0.31 (0.39)	−0.27 (0.40)	−0.34 (0.40)
322	3	−0.33 (0.42)	−0.37 (0.40)	−0.35 (0.40)	−0.34 (0.40)
134	1	0.66 (0.38)	0.10 (0.52)	−0.04 (0.41)	0.04 (0.41)
233	9	0.02 (0.40)	0.02 (0.40)	0.04 (0.41)	0.04 (0.41)
323	6	−0.05 (0.41)	−0.01 (0.41)	−0.04 (0.41)	0.04 (0.41)
332	4	0.18 (0.42)	0.10 (0.41)	0.11 (0.41)	0.04 (0.41)
234	1	0.69 (0.36)	0.32 (0.45)	0.36 (0.42)	0.44 (0.43)
243	1	0.60 (0.37)	0.45 (0.46)	0.52 (0.43)	0.44 (0.43)
333	24	0.37 (0.37)	0.44 (0.40)	0.44 (0.42)	0.44 (0.43)
342	2	0.83 (0.39)	0.58 (0.47)	0.60 (0.43)	0.44 (0.43)
423	1	0.54 (0.39)	0.34 (0.49)	0.36 (0.42)	0.44 (0.43)
441	1	1.31 (0.41)	1.11 (0.51)	0.68 (0.43)	0.44 (0.43)
334	5	0.89 (0.31)	0.79 (0.42)	0.78 (0.43)	0.86 (0.44)
343	1	0.82 (0.32)	0.90 (0.41)	0.95 (0.44)	0.86 (0.44)
442	1	1.40 (0.42)	1.15 (0.49)	1.04 (0.44)	0.86 (0.44)
344	4	1.25 (0.33)	1.30 (0.42)	1.32 (0.46)	1.32 (0.46)
434	2	1.24 (0.33)	1.25 (0.43)	1.23 (0.45)	1.32 (0.46)
443	1	1.17 (0.32)	1.37 (0.43)	1.41 (0.46)	1.32 (0.46)
444	7	1.81 (0.48)	1.88 (0.52)	1.88 (0.54)	1.88 (0.54)

overall patterns of the categorical response functions were similar in terms of plots, the continuation ratio model and the partial credit model yielded slightly different results from the graded response model and the generalized partial credit model. The model comparison using AIC indicated that the graded response model was the best fitting model to the data used in the illustration.

As long as the continuation ratio model yields similar item and ability parameters to other polytomous item response theory models as well as comparable information based goodness of fit measures, it can be viewed as an attractive alternative when polytomous items are analyzed. This study used a small data set for only a demonstration purpose. In order to understand the behavior of the item and ability parameter estimates under the continuation ratio model, a more extensive large scale simulation study should be performed.

It should be noted that in the continuation ratio model continuation ratio logits are sequentially used to model the probabilities of obtaining ordered categories in a polytomous item. In order to successfully apply the model to data, this sequential characteristic or nature of the assignment of ordered categories should be present in the construction of data. Inspecting the data if such a characteristic is present seems to be a prerequisite issue before applying logits to a multicategory variable.

In sum, the continuation ratio model considered in this paper can be applied to polytomous response items if they possess a special characteristic that the categories or ordered levels of the response are assigned in a forward, sequential manner. Note that not all polytomous, ordered responses have such a characteristic.

As long as the assumption is satisfied, the continuation ratio model is a unique model for the polytomous items due to the asymptotic independence of the categories within the item (cf. Fienberg 1980 pp. 110–111). Response categories of an item can be separately determined as if those were a set of dichotomous items. Hence, an application of the continuation ratio model in the context of differential item functioning may be promising because category response functions are rather independently obtained so that the category response functions from different groups can be directly compared (cf. Penfield, Gattamorta, & Childs, 2009). This model may also have a good potential use in metric linking and equating for polytomous items because the methods applicable to dichtomous items can be applied without any serious modifications (cf. Kim, Harris, & Kolen, 2010). The continuation ratio model may be a good choice for polytomous items when calibration is required for a test of items with mixed types (i.e., dichotomous and polytomous).

Appendix

```
FRENF.MLG
L2
>PROBLEM RANDOM, PATTERNS, NITEMS=9, NGROUPS=1, NPATTERNS=31,
 DATA='FRENF.DAT';
>TEST ALL, L2;
>END;
3
019
111111111
Y
9
(4X,9A1,F3.0)

111 099099099  4
```

```
112 099099109  1
121 099109099  4
211 109099099  1
122 099109109  4
212 109099109  1
221 109109099  4
123 099109110  2
132 099110109  1
222 109109109 10
312 110099109  1
223 109109110  5
232 109110109  8
322 110109109  3
134 099110111  1
233 109110110  9
323 110109110  6
332 110110109  4
234 109110111  1
243 109111110  1
333 110110110 24
342 110111109  2
423 111109110  1
441 111111099  1
334 110110111  5
343 110111110  1
442 111111109  1
344 110111111  4
434 111110111  2
443 111111110  1
444 111111111  7
```

References

Akaike, H. (1992). Information theory and an extension of the maximum likelihood principle. In S. Kotz, & N. L. Johnson (Eds.), *Breakthroughs in statistics: Vol. 1. Foundations and basic theory* (pp. 610–624). New York, NY: Springer. (Reprinted from *Second International Symposium on Information Theory,* pp. 267–281, by B. N. Petrov & F. Csaki, Eds., 1973, Budapest, Hungary: Akademiai Kiado).

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

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51. doi:10.1007/BF02291411.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459. doi:10.1007/BF02293801.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a micro-computer environment. *Applied Psychological Measurement, 6,* 431–444. doi:10.1177/014662168200600405.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B, 34,* 187–220.

du Toit, M. (Ed.). (2003). *IRT from SSI.* Lincolnwood, IL: Scientific Software International.

Fienberg, S. E. (1980). *The analysis of cross-classified categorical data* (2nd ed.). Cambridge, MA: The MIT Press.

Hambleton, R. K., van der Linden, W. J., & Wells, C. S. (2010). IRT models for the analysis of polytomously scored data: Brief and selected history of model building advances. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 21–42). New York, NY: Routledge.

Hemker, B. T., van der Ark, L. A., & Sijtsma, K. (2001). On measurement properties of continuation ratio models. *Psychometrika, 66,* 487–506. doi:10.1007/BF02296191.

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

Kim, S.-H. (2002, June). *A continuation ratio model for ordered category items.* Paper presented at the annual meeting of the Psychometric Society, Chapel Hill, NC. Retrieved from http://files.eric.ed.gov/fulltext/ED475828.pdf.

Kim, S.-H. (2013, April). *Parameter estimation of the continuation ratio model.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Kim, S., Harris, D. H., & Kolen, J. J. (2010). Equating with polytomous item response models. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory* (pp. 257–292). New York, NY: Routledge.

Luce, R. D., & Raiffa, H. (1957). *Games and decisions: Introduction and critical survey.* New York, NY: Wiley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174. doi:10.1007/BF02296272.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19,* 91–100. doi: 10.1177/014662169501900110.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176. doi: 10.1177/014662169201600206.

Penfield, R. D, Gattamorta, K., & Childs, R. A. (2009). An NCME instructional module on using differential step functioning to refine the analysis of DIF in polytomous items. *Educational Measurement: Issues and Practice, 28*(1), 38–49. doi: 10.1111/j.1745-3992.2009.01135.x.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph,* No. 17.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50,* 349–364.

Thissen, D., Chen, W.-H., & Bock, R. D. (2002). *MULTILOG: Multiple, categorical item analysis and test scoring using item response theory [Computer software].* Lincolnwood, IL: Scientific Software International.

Thissen, D., Nelson, L., Rosa, K., & McLeod, L. D. (2001). Item response theory for items scored in more than two categories. In D. Thissen, & H. Wainer (Eds.), *Test scoring* (pp. 141–186). Mahwah, NJ: Lawrence Erlbaum Associates.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology, 43,* 39–55. doi:10.1111/j.2044-8317.1990.tb00925.x

Using the Asymmetry of Item Characteristic Curves (ICCs) to Learn About Underlying Item Response Processes

Sora Lee and Daniel M. Bolt

Abstract In this chapter, we examine how the nature and number of underlying response subprocesses for a dichotomously scored item may manifest in the form of asymmetric item characteristic curves. In a simulation study, binary item response datasets based on four different item types were generated. The item types vary according to the nature (conjunctively versus disjunctively interacting) and number (1–5) of subprocesses. Molenaar's (2014) heteroscedastic latent trait model for dichotomously scored items was fit to the data. A separate set of simulation analyses considers also items generated with non-zero lower asymptotes. The simulation results illustrate that form of asymmetry has a meaningful relationship with the item response subprocesses. The relationship demonstrates how asymmetric models may provide a tool for learning more about the underlying response processes of test items. *online* at www.SpringerLink.com

Keywords Item response theory • Asymmetric ICCs • Item complexity • Item validity

1 Introduction

The item characteristic curves (ICCs) of most traditional item response theory (IRT) models are symmetric. Specifically, the change in probability observed above the inflection point in the ICC is a mirror image of the change that occurs below the inflection point. IRT models such as the Rasch model, the two and three-parameter logistic and normal ogive models are well-known examples.

Recently, there has been a growing psychometric literature related to asymmetric ICCs, and models that can be used to represent and explain such asymmetry. There are good reasons to believe that the nature of the psychological response process underlying many educational test items will be better reflected by asymmetric models. As considered by Samejima (2000), items scored as binary can often

S. Lee (✉) • D.M. Bolt
Department of Educational Psychology, University of Wisconsin, Madison, WI, USA
e-mail: slee486@wisc.edu; dmbolt@wisc.edu

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer
Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_2

be viewed as representing outcomes of multiple conjunctively or disjunctively interacting subprocesses. An example is a complex math word problem, in which the final answer may be arrived at only following the correct execution of a series of steps (e.g., converting the stated problem into an algebraic equation, solving the algebraic equation, etc.), where failure at any one step would lead to an overall incorrect response on the item. Assuming the individual steps (i.e., subprocesses) each conform to a logistic model, the overall item score should yield an asymmetric curve. In the case of conjunctively interacting subprocesses, the result should be an asymmetric ICC that accelerates at a slower rate to the right of the inflection point than it accelerates to the left of the inflection point (Samejima 2000). The extent of the asymmetry will be affected by the number of conjunctively interacting subprocesses.

Alternatively, for many items, the item score might be the outcome of disjunctively interacting subprocesses. An example is ability-based guessing model of San Martín, Del Pino, and De Boeck (2006), a model designed for multiple-choice items. Under the ability-based guessing model, a separate problem-solving process and guessing process are applied in sequential fashion such that an incorrect outcome from the problem solving process (e.g., the answer arrived at is not among the available response options), can be overcome by the guessing process. The nature of the asymmetry created by these two disjuctive subprocesses at the item score level (assuming again that each subprocess follows a logistic/normal ogive form) is the opposite to that described for the complex math word problem example. Specifically, the ICC will accelerate at a faster rate to the right of the inflection point than it accelerates to the left of the inflection point (Samejima 2000).

Model-based approaches to representing asymmetric ICCs of these kinds can take different forms. Samejima (2000) presents a logistic positive exponent (LPE) model in which an exponent parameter (or "acceleration" parameter) is introduced to a standard logistic model. While estimation algorithms have been proposed for this model (e.g., Samejima 2000; Bolfarine & Bazan, 2010), a challenge is the confound between the exponent parameter and the difficulty parameter (Lee 2015; Bolt, Deng, & Lee, 2014).

An alternative approach is Molenaar's (2014) normal ogive residual heteroscedasticity (RH) model. Molenaar (2014) illustrated how violation of the residual homoscedasticity assumption that underlies normal ogive models yields asymmetric ICCs for binary items. Such heteroscedasticity can be taken to reflect a greater variability in anticipated performances on an item conditional upon ability, and could conceivably reflect different underlying causes. In this chapter we consider the possibility that the heteroscedasticity reflects the nature and number of conjuctively/disjunctively interacting subprocesses described above, a feature that might often intuitively be expected to vary across items within a test. One of the advantages of the RH model is that the parameter associated with asymmetry is not confounded with difficulty, as in the LPE.

The purpose of this study is to examine whether the RH model can be used to inform about the underlying response processes associated with test items. Specifically, we examine how manipulation of both the nature and number of interacting subprocesses may be related to detectable asymmetries in the ICCs of

test items. Such an application, if successful, would support the RH model as item-level validation tool. From another perspective, it would suggest that the RH model may help in learning more about the underlying response process of a test item.

1.1 Other Implications of Ignoring Asymmetry in ICCs

The possibility that asymmetric ICCs can be used for item validation purposes represents just one additional reason for considering models such as the RH model.

The potential value of attending to asymmetry has already been considered from several different perspectives, suggesting that the implications of ignoring asymmetric ICCs, where they are present, can be significant. Woods and Harpole (2015), for example, have demonstrated the potential for inflated Type I error in DIF analyses when residual heteroscedasticity is present but ignored by the model testing for DIF. Molenaar (2014) illustrates how the estimated item information functions can be highly inaccurate when asymmetries are ignored. Such inaccuracies can not only influence how items are adaptively selected, but also the resulting estimated standard errors of ability estimates. With respect to person scoring, Samejima (2000) also notes an inconsistency in item weighting that emerges when using symmetric models, a problem that can be resolved using asymmetric models. Finally, ignoring asymmetry can also create problems related to the IRT metric. For example, Bolt et al. (2014) demonstrate how the presence of asymmetric ICCs may ultimately be responsible for the score deceleration problem seen when standardized tests are used to measure growth across grade levels.

1.2 Item Response Processes and Asymmetric ICCs

As indicated above, the purpose of this preliminary study was to examine whether the asymmetry of ICCs may also provide a way of learning about the nature and number of underlying item subprocesses, and whether the relationship is strong enough to allow asymmetric items to provide insight into the items. With multi-dimensional item response models, it has been common to attend to conjunctive or disjunctive response processes by considering different ways in which the latent traits, or more specifically, the processes associated with different latent traits, may interact. For example, cognitively diagnostic models emphasize skill attribute interactions as conjunctive versus disjunctive (e.g., Junker & Sijtsma, 2001; Maris 1995). Similarly, a distinction is often made between MIRT models that are compensatory versus noncompensatory (see e.g., Bolt & Lall, 2003). However, as emphasized in this paper, it can be useful to consider different forms of subprocess interaction in relation to collections of items that are statistically unidimensional. In Samejima's (2000) presentation of the LPE model, the number and nature of interacting subprocesses define the *complexity* of the item. We adopt the same terminology in this chapter, but use residual heteroscedasticity as a means of capturing such complexity as opposed to the exponent parameter used in the LPE.

2 Molenaar's Normal Ogive RH Model

The use of a normal ogive to represent an item response function for a binary item score follows from a model that assumes an underlying continuous latent response propensity that, conditional upon latent ability θ, is normally distributed. The mean of the conditional distribution is assumed to be a linear function of θ. The remaining variability in the response propensity conditional upon θ, denoted $\varepsilon_i|\theta$, represents sources of random noise, and is assumed to have a constant variance across θ, denoted $\sigma^2_{\varepsilon_i|\theta}$, referred to as the residual variance. In effect, scoring the item as binary can be viewed as defining a threshold with respect to ε_i that translates the continuous response propensity into a binary score. A normal ogive curve for the probability of correct response follows from the integration under the conditional normal distribution of the area above the threshold. Generalizations of this model to polytomous scores are straightforward, and simply require the consideration of multiple thresholds in relation to ε_i as opposed to just one (see e.g., Lord & Novick, 1968, pp. 370–371 for details).

The assumption of homoscedasticity of the response propensity variance across ability levels naturally plays an important role in how the probability of a correct response is defined. If heteroscedasticity of variance is present, it will alter the form of the probability curve assuming other features of the model are held constant. Generalization of the normal ogive model to accommodate heteroscedasticity of variance naturally requires specification of a suitable function for $\sigma^2_{\varepsilon_i|\theta}$. Molenaar proposed the following form of heteroscedasticity in the context of polytomously scored items (Molenaar, Dolan, & De Boeck, 2012):

$$\sigma^2_{\epsilon_i|\theta} = 2\delta_0[1 + \exp(-\delta_1\theta)]^{-1} \tag{1}$$

where δ_0 is a baseline parameter, and δ_1 is heteroscedasticity parameter, $\delta_0 \in (0, \infty)$ and $\delta_1 \in (-\infty, \infty)$. Note that if $\delta_1 = 0$, then the residual variances are homoscedastic with $\sigma^2_{\varepsilon_i|\theta} = \delta_0$; if $\delta_1 > 0$, then the residual variance is increasing with θ; if $\delta_1 < 0$, residual variances are decreasing with θ.

Molenaar (2014) derived a corresponding model for dichotomously scored items based on the same model for heteroscedasticity. The resulting item response function is:

$$P(y_i = 1|\theta) = \Phi\left(\frac{\alpha_i\theta + \beta_i}{\sqrt{2}[1 + \exp(-\delta_{1i}\theta)]^{-1/2}}\right) \tag{2}$$

where δ_{1i} is the item heteroscedastic parameter, and α_i and β_i denote the slope (discrimination) and intercept (difficulty) parameters associated with the normal ogive model. As for the polytomous model, the model in (2) reduces to the standard normal ogive model in the case where $\delta_{1i} = 0$. Further details on this model are provided by Molenaar (2014).

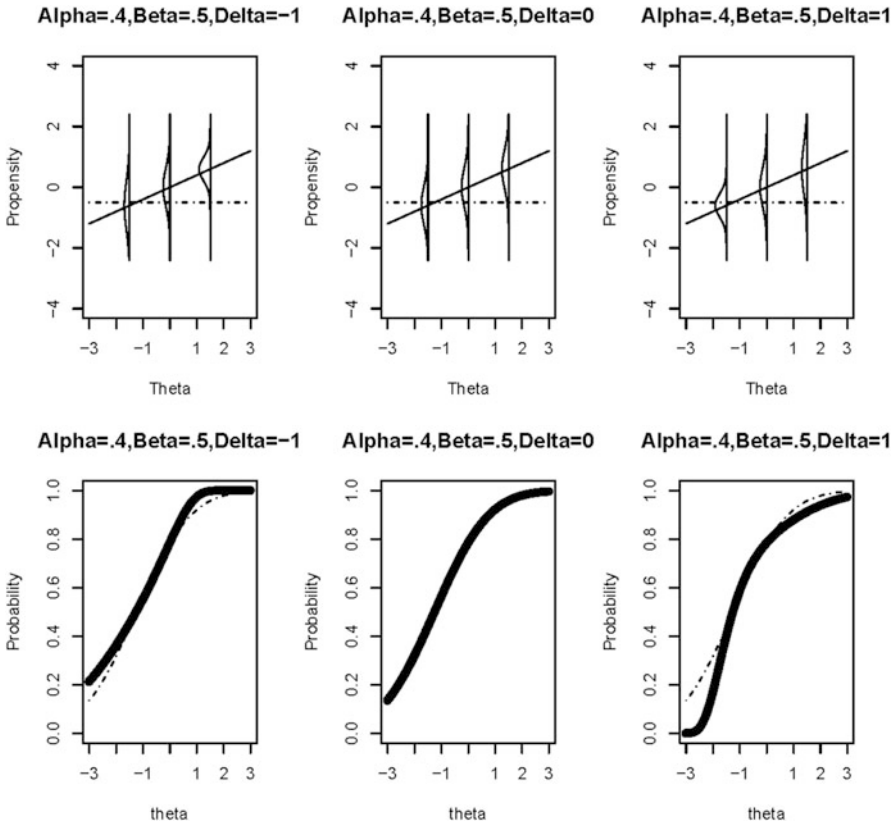


Fig. 1 Residual heteroscedasticity and item characteristic curves

Figure 1 provides an illustration of how manipulation of the δ_{1i} parameter introduces ICC asymmetry. The plots at the top of the figure illustrate the heteroscedasticity associated with the RH model for three different hypothetical items that vary only with respect to δ_{1i}. The middle figure corresponds to the condition of homoscedasticity, while the figures on the left and right correspond to examples where the residual variance decreases and increases, respectively, in relation to θ. When translated into probability curves, the items yield different ICCs. In particular, a negative δ_{1i} results in an ICC with a steeper slope to the right of the inflection point than the corresponding symmetric ICC, and a flatter slope to the left of the inflection point. Just the opposite is observed for the item with a positive delta value.

A primary goal of the current paper is to illustrate how Molenaar's RH model can be used to capture differences in the underlying response processes associated with different items. To this end, we also attempt to illustrate how asymmetric ICCs can be a naturally expected outcome for educational test data. We also seek to clarify the potential of the RH model in recovering the nature of the asymmetry associated with these different response processes.

It is worth noting that estimation procedures also exist for other models that can flexibly account for asymmetric ICCs. For example, Bolfarine and Bazan (2010) considered the use of Bayesian estimation techniques with Samejima's LPE model. Preliminary work (Lee 2015), however suggests that the RH model of Molenaar may be slightly better in terms of recovery, perhaps in large part due to the greater separation of parameters associated with the asymmetry and item difficulty. We therefore focus on Molenaar's RH model in the current paper.

2.1 *Bayesian Estimation of Heteroscedastic Two-Parameter and Three-Parameter Normal Ogive Models*

Molenaar (2014) presents a marginal maximum likelihood algorithm for the RH model. In this paper we consider the model in a Bayesian estimation framework, as well as a three-parameter version that introduces a lower asymptote parameter.

Under the two-parameter Residual Heteroscedasticity (2P-RH) model, we assume the following priors for the item parameters:

$$\beta_i \sim \text{Normal}(0,1)$$

$$\alpha_i \sim \text{Lognormal}(0,2)$$

$$\delta_{1i} \sim \text{Normal}(0,1)$$

and for the person parameter:

$$\theta \sim \text{Normal}(0,1)$$

For the three-parameter Residual Heteroscedasticity (3P-RH) model, we consider use of the same parameters, but add a fixed lower asymptote parameter, γ:

$$P(y_i = 1|\theta) = \gamma + (1-\gamma)\Phi\left(\frac{\alpha_i\theta + \beta_i}{\sqrt{2}[1+\exp(-\delta_{1i}\theta)]^{-1/2}}\right) \tag{3}$$

In the current study, $\gamma = 0.2$ when generating the data, and we also fix $\gamma = 0.2$ when estimating the model, as might reflect a multiple-choice test with five options per item. Thus the three-parameter simulation evaluates how well the model functions in the presence of known guessing effects. Our preliminary analyses did consider a 3P-RH model with an estimated lower asymptote, although the model resulted in simulated chains with poor convergence.

3 Simulation Study

To evaluate the effectiveness of the RH model in informing about underlying response process, we simulated item response data to conform to different types of response processes. In effect, we assumed each binary item was the outcome of one of four possible types, ordered from the least to most complex: (1) a disjunctive two-subprocess item; (2) a single subprocess item; (3) a conjunctive two-subprocess item; and (4) a conjunctive five-subprocess item. In all cases, data were simulated as unidimensional. It is worth noting that unlike models such as in Whitely (1980), the presence of distinct subprocesses is not associated with multidimensionality, reflecting the fact that as a statistical dimension, a single underlying latent trait can often reflect what is in reality a complex constellation of skills. Regardless of the item type, each subprocess was simulated from a normal ogive model, i.e.,

$$P_{ik}(\theta) = P(u_{ik} = 1|\theta) = \Phi(\alpha_{ik}\theta + \beta_{ik}), \tag{4}$$

where $P_{ik}(\theta)$ denotes the probability of successfully executing subprocess k on item i (i.e., $u_{ik} = 1$), and α_{ik}, β_{ik} denote item subprocess discrimination and difficulty (threshold) parameters, respectively. The distinguishing characteristics of the items relate to the number of subprocesses as well as the nature of their interaction.

3.1 Low Complexity Disjunctive Items: A Two Subprocess Model

The first item type simulated assumes two subprocesses with a disjunctive interaction:

$$P(y_i = 1|\theta) = P_{i1}(\theta) + (1 - P_{i1}(\theta))P_{i2}(\theta) \tag{5}$$

As noted earlier, such a model could reflect an ability-based guessing context (San Martín et al. 2006), whereby a student can solve the item in one of two ways: (1) ordinary problem solving behavior, where the solution may be arrived at using the intended approach, while if not attained is followed by (2) guessing behavior, where the various response options are evaluated apart from the intended problem-solving process, and the most sensible option is chosen.

3.2 Moderate Complexity Items: One Subprocess Model

For comparison purposes, we consider also a one subprocess item:

$$P(y_i = 1|\theta) = P_{i1}(\theta) \tag{6}$$

Such items reflect a condition of ICC symmetry, and might correspond to items that reflect direct testing of particular components of knowledge, such as the definition of a concept, for example. With respect to the RH model, they should yield items for which the estimated δ_{1i} is near 0.

3.3 Moderately High Complexity Conjunctive Items: A Two Subprocess Model

A third item type assumes two subprocesses, but with a conjunctive interaction:

$$P(y_i = 1|\theta) = P_{i1}(\theta)P_{i2}(\theta) \tag{7}$$

From the section above, it is anticipated that conjunctive items will yield positive delta estimates. Such items would represent an item that involves two steps, where a correct answer is only achieved when both steps are successfully executed.

3.4 High Complexity Conjunctive Items: A Five Subprocess Model

The fourth item type is similar to the third, but involves five, as opposed to two, subprocesses:

$$P(y_i = 1|\theta) = P_{i1}(\theta)P_{i2}(\theta)P_{i3}(\theta)P_{i4}(\theta)P_{i5}(\theta) \tag{8}$$

Such items could be viewed as items involving five steps, where a correct answer is only attained when all five steps are executed correctly. Relative to the previous category, these items should return the most positive estimates of δ_{1i}.

In order to simulate items that varied primarily in the number and nature of interacting subprocesses, we simulated subprocess parameters using distributions within each item type that would render items that were comparable in terms of overall item discrimination and difficulty. It was our intent that the primary psychometric feature distinguishing these four categories of item types from each other would be the asymmetry of their ICCs, not characteristics such as difficulty or discrimination. For the five subprocess conjunctive items, we generated $\alpha_{ik} \sim$ lognorm(-0.3,0.4) and $\beta_{ik} \sim$ unif(1,2.5); for the two subprocess conjunctive items, $\alpha_{ik} \sim$ lognorm(-0.1,0.4) and $\beta_{ik} \sim$ unif(0,1.5); for the two subprocess disjunctive model, $\alpha_{ik} \sim$ lognorm(0,0.4) and $\beta_{ik} \sim$ unif(-1.5,1); for the one subprocess model, $\alpha_{i1} \sim$ lognorm(0,0.4) and $\beta_{i1} \sim$ unif(-2,2).
In all cases, we also simulated examinee proficiency θ as normal, with a mean of 0 and variance of 1.

Finally, in order to consider a situation in which a non-zero lower asymptote was also present, in a separate set of simulation analyses, we generated items from the same four item type categories but now using a simulation model that introduced a nonzero lower asymptote. Specifically, for the low complexity disjunctive items we simulate:

$$P(y_i = 1|\theta) = \gamma + (1 - \gamma)[P_{i1}(\theta) + (1 - P_{i1}(\theta))P_{i2}(\theta)],$$

while for the moderate complexity items:

$$P(y_i = 1|\theta) = \gamma + (1 - \gamma)P_{i1}(\theta),$$

for the moderately high complexity conjunctive items:

$$P(y_i = 1|\theta) = \gamma + (1 - \gamma)P_{i1}(\theta)P_{i2}(\theta),$$

and for the high complexity conjunctive items:

$$P(y_i = 1|\theta) = \gamma + (1 - \gamma)P_{i1}(\theta)P_{i2}(\theta)P_{i3}(\theta)P_{i4}(\theta)P_{i5}(\theta),$$

where in all cases, $\gamma = .2$. As described above, we also fixed the γ at .2 when estimating the model.

Each simulated dataset included ten items from each category, so 40 items total per simulated dataset, and simulated responses for 25,000 examinees. All MCMC runs were run out to 10,000 iterations, and δ_{1i} estimates were obtained for each item. We carried out 20 replications for each of the two-parameter and three-parameter simulation models. In each case the appropriate model (two-parameter or three-parameter RH model) was used as corresponded to the simulation condition.

4 Simulation Results

Figure 2 provides a graphical illustration of the δ_{1i} estimates for a single simulation run in each of the two-parameter and three-parameter conditions against the item type category. The item type categories are ordered from least to most complex, such that the increase in δ_{1i} estimates across categories is as expected. Tables 1 and 2 provide a tabulation of the results across 20 replications in each condition. Also apparent from the table is the tendency for the δ_{1i} estimates to increase as item complexity increases. Nevertheless, there remains a fair amount of variability within each category, variability that can be attributed to the imprecision in estimating δ_{1i} as well as the potential sensitivity of the δ_{1i} estimates to other characteristics of items (e.g., the difficulty and discrimination of the individual subprocesses within item) that varied within the simulation and may have an effect on these estimates. It is, however, noteworthy that the vast majority of items in the low complexity

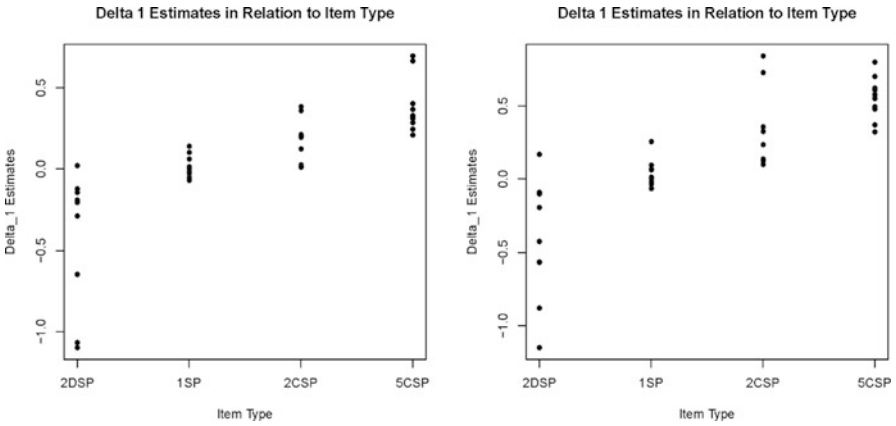


Fig. 2 δ_{1i} estimates against the item type category in 2P (*left*) and 3P (*right*) condition, respectively

Table 1 δ_{1i} estimates against the item type in 2P condition (ICC = 0.65)

Item type	$\hat{\delta}_1$ Mean	$\hat{\delta}_1$ Std dev
2DSP	−0.39	0.41
1SP	0.01	0.07
2CSP	0.17	0.13
5CSP	0.38	0.17

Table 2 δ_{1i} estimates against the item type in 3P condition (ICC = 0.64)

Item type	$\hat{\delta}_1$ Mean	$\hat{\delta}_1$ Std dev
2DSP	−0.39	0.41
1SP	0.04	0.09
2CSP	0.31	0.27
5CSP	0.55	0.14

category return δ_{1i} estimates less than 0, while those in the moderate complexity category are centered right around 0, and the vast majority of those in the moderate or high complexity category return δ_{1i} estimates greater than 0. Intraclass correlation estimates, which are from variance component estimation using the ANOVA method to determine within and between item type variance, were 0.65 and 0.64 for the two-parameter and three-parameter analyses, respectively, suggesting that the presence of a nonzero lower asymptote (corresponding to the effects of random guessing) does not have a deleterious effect on the δ_{1i} estimates. It is also worth noting, however, that the category of low item complexity seemed to yield the highest variability in δ_{1i} estimates. Such a result may reflect the metric of the δ_{1i} parameter.

5 Discussion

There are several limitations to our study. First, it is only a simulation, and should be replicated with real data. Identifying example items where the underlying response process is known or highly suspected, and seeing δ_{1i} estimates from real data analyses that are consistent with such knowledge, would provide strong evidence in support of the approach. Second, our simulation used a proficiency distribution that matched that assumed by the estimation algorithm (in both cases normal). The possibility of non-normal trait distributions, and the implications this has for representing asymmetries and how they vary across items, should be further examined. The shape of any ICC is to a large extent arbitrary when considering arbitrary nonlinear alterations of the proficiency metric. Alternative approaches have considered retaining the symmetric model, but allowing for nonnormal trait distributions (see e.g., Woods & Thissen, 2006). The possibility of altering the ICC shape versus altering the proficiency metric is often unclear when analyzing real data (Molenaar 2014). The presence of items that vary in the number and nature of subprocesses is important in generating meaningful variability in delta. Third, the nature of the response processes for the different item type categories are simplistic. It is of course conceivable that an item may contain a mix of conjunctively and disjunctively interacting subprocesses, and that many items may also be solved using multiple different strategies. Fourth, our simulation study used large samples, as may often be available for large-scale assessments. It remains to be seen how well the model performs with smaller samples.

There are also additional extensions to the method and its application that could be considered. As noted earlier, the possibility of estimating a lower asymptote parameter for the RH model could be considered. In addition, other forms of heterscedasticity in relation to the proficiency could be developed, some of which may be more appropriate than the current approach for the types of items being simulated. In general, beyond seeing relationships between the δ_{1i} parameter and item type category, more work is needed in evaluating how well the RH model actually fits items of the type simulated in this chapter. Finally, the possibility of using the RH model as a basis for IRT applications, such as CAT or vertical scaling, and comparisons against traditional approaches using symmetric models, would be useful.

References

Bolfarine, H., & Bazan, J. L. (2010). Bayesian estimation of the logistic positive exponent IRT model. *Journal of Educational and Behavioral Statistics, 35*, 693–713.

Bolt, D. M., Deng, S., & Lee, S. (2014). IRT model misspecification and measurement of growth in vertical scaling. *Journal of Educational Measurement, 51*(2), 141–162.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional IRT models using Markov chain Monte Carlo. *Applied Psychological Measurement, 27*, 395–414.

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.

Lee, S. (2015). *A comparison of methods for recovery of asymmetric item characteristic curves in item response theory* (Unpublished master's thesis). Madison: University of Wisconsin.

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

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60*, 523–547.

Molenaar, D. (2014). Heteroscedastic latent trait models for dichotomous data. *Psychometrika*, 80(3), 625–644.

Molenaar, D., Dolan, C. V., & De Boeck, P. (2012). The heteroscedastic graded response model with a skewed latent trait: Testing statistical and substantive hypotheses related to skewed item category functions. *Psychometrika, 77*, 455–478.

Samejima, F. (1995). Acceleration model in the heterogeneous case of the general graded response model. *Psychometrika, 60*(4), 549–572.

Samejima, F. (2000). Logistic positive exponent family of models: Virtue of asymmetric item characteristic curves. *Psychometrika, 65*, 319–335.

San Martín, E., Del Pino, G., & De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement, 30*(3), 183–203.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*(4), 479–494.

Woods, C. M., & Harpole, J. K. (2015). How item residual heterogeneity affects tests for differential item functioning. *Applied Psychological Measurement, 39*, 251–263.

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

A Three-Parameter Speeded Item Response Model: Estimation and Application

Joyce Chang, Henghsiu Tsai, Ya-Hui Su, and Edward M. H. Lin

Abstract When given time constraints, it is possible that examinees leave the harder items till later and are not able to finish answering every item in time. In this paper, this situation is modeled by incorporating a speeded-effect term into a three-parameter logistic item response model. Due to the complexity of the likelihood structure, a Bayesian estimation procedure with Markov chain Monte Carlo method is presented. The methodology is applied to physics examination data of the Department Required Test for college entrance in Taiwan for illustration.

Keywords Item response model • Markov chain Monte Carlo • Test speededness

1 Introduction

Over the past few decades, there has been increasing interest in modeling response data generated from tests that are administered within an allocated time, which may be insufficient for some examinees. A test is said to be speeded if the time limit affects examinees' test performance (see, for example, Lee & Ying 2015). In order to reduce the contamination of the test speededness in modeling response

J. Chang
Department of Economics, The University of Texas at Austin, 2225 Speedway, BRB 1.116, C3100, Austin, Texas 78712, USA
e-mail: joyce.chang@utexas.edu

H. Tsai
Institute of Statistical Science, Academia Sinica, 128 Academia Road, Section 2, Nangang District, Taipei 11529, Taiwan
e-mail: htsai@stat.sinica.edu.tw

Y.-H. Su (✉)
Department of Psychology, National Chung Cheng University, 168 University Road, Section 1, Min-Hsiung, Chai-Yi 62102, Taiwan
e-mail: psyyhs@ccu.edu.tw

E.M.H. Lin
Institute of Finance, National Chiao Tung University, 1001 University Road, Hsinchu 300, Taiwan
e-mail: m9281067@gmail.com

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_3

data, several models have been proposed in the literature. Yamamoto (1995) uses the HYBRID model to describe the behavior that an examinee may switch to a guessing strategy midway through a test due to the time constraint. Unlike the unspeeded items, which are characterized by a two-parameter logistic (2PL) model, the speeded ones are, on the other hand, characterized by a latent class based item response model. Bolt, Cohen, and Wollack (2002) use the mixture Rasch model of Rost (1990) to deal with situations where no penalty is imposed for guessing; consequently, speededness effects tend to emerge in the form of incorrect as opposed to omitted responses. Goegebeur, De Boeck, Wollack, and Cohen (2008) propose a speeded item response theory (IRT) model with gradual process change. Under this model, responses to items early in the test are governed by a 3PL model, and beyond some point the success probability gradually decreases and eventually reduces to the success probability under random guessing. Chang, Tsai, and Hsu (2014) propose the leave-the-harder-till-later speeded two-parameter logistic (LHL-2PL) model to accommodate the speeded effect. Additional literature on test speededness includes Bejar (1985), Yamamoto (1989), Yamamoto and Everson (1997), Boughton and Yamamoto (2007), Cao and Stokes (2008), and Wang and Xu (2015), among others.

In this paper, we are interested in extending the LHL-2PL model by adding a pseudo-guessing parameter. Chang, Tsai, and Hsu (2014) apply the LHL-2PL model to the physics examination data of Department Required Test (DRT) for college entrance in Taiwan, and find some evidence for the LHL mechanism in analyzing the data. Examinees have to answer 26 questions in 80 min, where the first 20 questions are multiple-choice questions that examinees should choose one correct answer out of 5 possible choices. It is then followed by 4 multiple-response questions, where out of the 5 possible, examinees need to select all the answer choices that apply, and finally 2 calculation problems. The test is administered under formula-scoring directions, where 3/4 and 1 point are deducted from the raw score for each incorrect answer made in the multiple-choice and multiple-response questions respectively. If an item is left blank, the examinee would get 0 point. Furthermore, the adjusted score would only be 0 or above for these two types of questions.

Based on the discussions of Lord (1975) on formula scoring, Chang, Tsai, and Hsu (2014) argue that examinees are less likely to guess whenever they do not know the answer, and therefore, it provides some rationale for considering a speeded model in which random guessing is not allowed. However, it is also argued that examinees often know enough about the subject to eliminate some of the incorrect choices. That being the case, guessing from among the remaining options is likely to help them overcome the penalty of $1/(k-1)$, where k is the number of options, and is 5 for the first 20 multiple-choice questions (e.g., Angoff 1989). For each of the 4 multiple-response questions, there are 5 choices, and each one is graded independently, so $\mathrm{k} = 2$. That is, each choice in the multiple-response question is either true or false. In the literature, many papers also allow random guessing (or pseudo-guessing) parameters in their models, see, for example, Cao and Stokes (2008), Goegebeur, De Boeck, Wollack, and Cohen (2008), and Wang and Xu (2015). This motivates us to consider in this paper the leave-the-

harder-till-later speeded three-parameter logistic IRT (LHL-3PL) model by adding a pseudo-guessing parameter to the LHL-2PL model of Chang, Tsai, and Hsu (2014).

The rest of the paper is organized as follows. In Sect. 2, we describe the LHL-3PL model in more details. Since our model is a direct extension of Chang, Tsai, and Hsu (2014), our prior settings are the same as theirs except for the extra pseudo-guessing parameters. The prior settings for the pseudo-guessing parameters will also be mentioned in Sect. 2. A simulation study is conducted in Sect. 3 to demonstrate the validation of the Bayesian estimation procedure. Application of the LHL-3PL model to the data of Department Required Test for college entrance in Taiwan is illustrated in Sect. 4. Section 5 concludes.

2 Leave-the-Harder-till-Later Speeded Three-Parameter Logistic Item Response Model

Let Y_{pj} be the dichotomous response of examinee p on item j, where $p = 1, 2, \ldots, P$, and $J = 1, 2, \ldots, J$. Denote b_j and a_j as the location and scale parameters respectively, for item j, and θ_p as the ability parameter for examinee p. In the 2PL model (Birnbaum 1968), the probability that examinee p gets a correct response on item j is given by

$$\Pr(Y_{pj} = 1|a_j, b_j, \theta_p) = \frac{1}{1 + e^{-a_j(\theta_p - b_j)}}.$$

The parameter a_j is also known as the discrimination parameter (de Ayala 2009), or the slope parameter (Wang 2004), and the parameter b_j is called the difficulty parameter in Embretson and Reise (2000) and Wang and Xu (2015). For more descriptions and discussions of the 2PL model, see Embretson and Reise (2000), Wang (2004), and de Ayala (2009).

The three-parameter logistic (3PL) model is obtained by adding an extra parameter to the 2PL model. Under the 3PL model,

$$\Pr(Y_{pj} = 1|a_j, b_j, c_j, \theta_p) = c_j + (1 - c_j) \cdot \frac{1}{1 + e^{-a_j(\theta_p - b_j)}}.$$

The parameter c_j is referred to as the item's pseudo-guessing or pseudo-chance parameter and equals the probability of a correct response when θ approaches $-\infty$ (de Ayala 2009). It is also named the asymptotic parameter (Wang 2004) or the lower-asymptotic parameter (Embretson & Reise 2000). The 3PL model is suitable for multiple-choice cognitive items (Embretson & Reise 2000; Wang 2004).

Unlike the traditional IRT models described above, where unspeededness is implicitly assumed, Chang, Tsai, and Hsu (2014) introduce two additional parameters to the 2PL model in an attempt to capture the effect of speededness. It is assumed that the probability of a correct response in given by

$$\Pr\left(Y_{pj}=1\left|a_j,b_j,\theta_p,\tau_p,\lambda\right.\right)=\frac{e^{-\lambda\left(b_j-\tau_p\right)\cdot\mathrm{I}\left\{b_j>\tau_p\right\}}}{1+e^{-a_j\left(\theta_p-b_j\right)}},\tag{1}$$

where τ_p is the p-th examinee's threshold parameter for speededness and λ, which is always larger than zero, is the speededness rate. Indicator function $I\{\cdot\}$ is defined as

$$I\{b_j>\tau_p\}=\begin{cases}1, & b_j>\tau_p,\\ 0, & b_j\le\tau_p.\end{cases}$$

The rationality behind the model is as follows. When encountering an item, the examinee would decide if he would get into solving process right away by the level of difficulty of the item. If its difficulty exceeds one's threshold, τ_p, i.e., $b_j>\tau_p$, the item is considered time-consuming and would be retained till a later test period. It is further assumed that the first-skipped item would be answered with the probability of $e^{-\lambda(b_j-\tau_p)}$. In other words, the model can be partitioned into two parts: (1) whether to solve or not, and (2) whether the answer is correct. The two stages are given by

$$Z_{pj}|(b_j,\tau_p,\lambda)\sim\text{Bernoulli}\left(e^{-\lambda(b_j-\tau_p)\cdot I\{b_j>\tau_p\}}\right),$$

$$Y_{pj}|(a_j,b_j,\theta_p,Z_{pj})\sim\text{Bernoulli}\left(\frac{1}{1+e^{-a_j\left(\theta_p-b_j\right)}}\cdot Z_{pj}\right),$$

where Z_{pj} denotes whether the item is being answered or not.

As discussed in Sect. 1, for the DRT data, the first 20 questions and the 21st to the 24th questions are multiple-choice questions and multiple-response questions respectively, and are therefore, naturally suitable for a 3PL model, where a pseudo-guessing parameter is included. Specifically, we consider the LHL-3PL model (to be defined below). For the last 2 calculation problems, we simply set the corresponding pseudo-guessing parameters to be zero. Under the LHL-3PL model,

$$\Pr\left(Y_{pj}=1\left|a_j,b_j,c_j,\theta_p,\tau_p,\lambda\right.\right)=c_j+\left(1-c_j\right)\cdot\frac{e^{-\lambda\left(b_j-\tau_p\right)\cdot\mathrm{I}\left\{b_j>\tau_p\right\}}}{1+e^{-a_j\left(\theta_p-b_j\right)}},\tag{2}$$

where $0<c_j<1$. We want to compare our proposed LHL-3PL model with the LHL-2PL of Chang, Tsai, and Hsu (2014) to explore the role of random guessing in the DRT data, so we adopt the assumptions, including the normality of the joint distribution of θ_p and τ_p, prior settings and the MCMC-based estimation procedure of Chang, Tsai, and Hsu (2014). For the pseudo-guessing parameter c_j, we transform it into the real number scale γ_j, and assume

$$\gamma_j=\log\left(\frac{c_j}{1-c_j}\right)\sim\mathrm{N}\left(\mu_\gamma,\sigma_\gamma^2\right),\tag{3}$$

Table 1 RMSE of estimates from LHL-3PL fitting under data generated from the LHL-3PL model (10 replicates)

Parameter\ P	250	500	1,000
$\boldsymbol{b}$	0.9521	0.9392	0.7881
$\boldsymbol{a}$	1.4735	0.8152	0.7369
$\boldsymbol{c}$	0.0897	0.0978	0.0978
$\boldsymbol{\theta}$	0.5645	0.5387	0.5306
$\boldsymbol{\tau}$	2.8719	2.8198	2.7675

and

$$\mu_\gamma \sim \mathrm{N}\left(\mu, \sigma^2\right), \qquad \sigma_\gamma^2 \sim \mathrm{Inv-Gamma}\left(\alpha, \beta\right), \tag{4}$$

where $\mu = 0, \sigma^2 = 1, \alpha = \beta = 3$.

Bayesian estimation method has been widely used in IRT modeling, see, for example, Swaminathan and Gifford (1982, 1985, 1986), Mislevy (1986), Bolt, Cohen, and Wollack (2002), van der Linden (2007), Cao and Stokes (2008), Fox (2010), Meyer (2010), and Chang, Tsai, and Hsu (2014).

3 Simulation Study

In this section, we conduct a simulation study to evaluate the performance of the MCMC method in estimating the parameters. All computations were performed using some Fortran code with IMSL subroutines.

We first describe the true data generating process. We consider $J = 40, P = 250$, 500, and 1,000. Let $\boldsymbol{a} = (a_1, \cdots, a_J)$, $\boldsymbol{b} = (b_1, \cdots, b_J)$, $\boldsymbol{c} = (c_1, \cdots, c_J)$, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_P)$, and $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_P)$. The true values of $\boldsymbol{a}$ and $\boldsymbol{b}$ are the same as those considered in Sect. 4 of Chang, Tsai, and Hsu (2014). For the true values of $\boldsymbol{c}$, we set $c_j = (40.5 - j)/40$, for $j = 1, \ldots, 40$. The true value of λ equals 1. For $p = 1, \ldots, P$, (θ_p, τ_p) are independently and identically sampled from a bivariate normal distribution with the marginal distribution of θ_p and τ_p being $N(0, 1)$ and $N(0.2, 0.5)$, respectively, and the correlation being 0.8.

We produce 40,000 MCMC draws with the first 10,000 draws as burn-in. For each parameter, the posterior mean was calculated as our Bayes estimates, based on 30,000 MCMC draws after burn-in. We repeat the exercise 10 times, and the root mean squared error (RMSE) of the posterior means are summarized in Table 1. From Table 1, it is clear that, in general, the RMSE decreases with the value P, except for the parameter $\boldsymbol{c}$. However, the RMSE's of the parameter $\boldsymbol{c}$ are the smallest, and those of the parameter $\boldsymbol{\tau}$ are the largest. From $P = 250$ to $P = 1{,}000$, the RMSE's of the parameter $\boldsymbol{a}$ become half.

4 Application

In this section, the proposed LHL-3PL model and the MCMC procedure described in the previous section are applied to the data of the physics examination of the 2010 Department Required Test for college entrance in Taiwan provided by College Entrance Examination Center (CEEC). The data from 1,000 randomly sampled examinees contains the original responses and nonresponses information, but we treat both nonresponses and incorrect answers the same way and code them as $Y_{pj} = 0$ as suggested by Chang, Tsai, and Hsu (2014). As for the calculation part, the response Y_{pj} is coded as 1 whenever the original score is more than 7.5 out of 10 points, and zero otherwise.

The four models, including the 2PL, LHL-2P, 3PL, and the LHL-3PL models, are fitted to the data using Bayesian analysis. For the 3PL and the LHL-3PL models, we set $c_{25} = c_{26} = 0$ because guessing is in theory not possible. Further comparison is made via Bayesian model selection criterion, the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde 2002), described below.

We use the posterior means as the point estimates for parameters of interest. Let $\boldsymbol{\xi} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\theta}, \boldsymbol{\tau}, \lambda)$, and $\hat{\boldsymbol{\xi}} = (\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{c}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\tau}}, \hat{\lambda})$ be the posterior mean of $\boldsymbol{\xi}$ under the fitted LHL-3PL model given data $\mathbf{y} = (\mathbf{y_1}, \cdots, \mathbf{y_P})$, where $\mathbf{y_p} = (\mathrm{y}_{p1}, \cdots, \mathrm{y}_{pJ})$. The DIC for the fitted LHL-3PL model is defined as

$$\mathrm{DIC} = D(\hat{\boldsymbol{\xi}}) + 2p_D, \tag{5}$$

where

$$D(\hat{\boldsymbol{\xi}}) = -2\log f(\mathbf{y}|\hat{\boldsymbol{\xi}}),$$
$$p_D = E_{\boldsymbol{\xi}|\mathbf{y}}[-2\log f(\mathbf{y}|\boldsymbol{\xi})] - D(\hat{\boldsymbol{\xi}}).$$

In (5), the first term $D(\hat{\boldsymbol{\xi}})$ measures the goodness-of-fit, and the second term p_D, which represents the effective number of parameters used in the model, is the difference between posterior mean deviance and deviance evaluated at the posterior means of the parameters. The DIC for the other three fitted models are defined similarly. A smaller DIC is preferred, which selects a model with a better goodness-of-fit and simultaneously maintains the model complexity to be as simple as possible. The resulting DIC values for the four fitted models are listed in the second row of Table 2. The LHL-3PL has a smallest DIC, indicating the best fitting performance of the LHL-3PL as compared to the other models after compensating for model complexity.

Apart from DIC, the Bayesian model-data fit checking techniques, such as posterior predictive model checking (PPMC), has also been used in the literature. See, for example, Li, Bolt, and Fu (2006), Sinharay, Johnson, and Stern (2006), and Huang and Hung (2010). The procedure runs as follows:

Table 2 DIC for physics examination data of the Department Required Test for college entrance in Taiwan

Model	2PL	LHL-2PL	3PL	LHL-3PL
DIC	24,671.99	24,717.57	24,506.24	24,416.17

Step 1. Compute the realized discrepancy measure from the observed data set $\mathbf{y}$.
Step 2. Generate a draw of parameter $\boldsymbol{\xi}$ from the posterior distribution.
Step 3. Draw a data set $\tilde{\mathbf{y}}$ from the model, using the parameter $\boldsymbol{\xi}$ drawn in Step 2.
Step 4. Compute the value of the predictive discrepancy measure from the above draws of parameters and data set $\tilde{\mathbf{y}}$.
Step 5. Repeat Steps 2–4 1,000 times to compute the posterior predictive p-value (PPP-value).

The PPP-value is defined to be the percent of times that the predictive discrepancy measure is larger than its realized counterpart. An extreme PPP-value (PPP-value larger than 0.975 or smaller than 0.025) suggests that the model fits the data poor (Li, Bolt, & Fu 2006, p. 11). Following from Li, Bolt, and Fu (2006) and Sinharay, Johnson, and Stern (2006), we use the sample odds ratio (e.g. Agresti 2002p. 45) as the discrepancy measure in our study. The sample odds ratio is defined to be $OR = (n_{11}n_{00})/(n_{10}n_{01})$, where n_{jk} denotes the number of individuals scoring j on the first item and k on the second item, $j, k = 0, 1$. The sample odds ratio tests item response association between a pair of items. Here, we have $J = 26$ items, resulting in $J(J-1)/2 = 325$ pairs, and therefore, 325 PPP-values. The number of extreme PPP-values of the four fitted models are all zeros, indicating the goodness of fits of these four models.

Let $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_J)$, where, for $j = 1, \ldots, J$, $\pi_j = \sum_{p=1}^{P} y_{pj}/P$. Thus, for $j = 1, \ldots, 24$, π_j represents the percent of examinees who respond correctly to question j, and for $j = 25$ and 26, it represents the percent of examinees whose original score is more than 7.5.

Now, we compare the estimates of these four models. Since the estimates of 2PL and LHL-2PL are similar, and those of 3PL and LHL-3PL are similar, we only compare those of LHL-2PL and LHL-3PL in the following. Figure 1a shows the plots of $\hat{c}_j$ and π_j, over $j = 1, \ldots, 26$. Recall that $c_{25} = c_{26} = 0$. From Fig. 1a, we see that fewer examinees score more than 7.5 or above in the calculation problems than getting a correct answer on each of the multiple-choice questions or the multiple-response questions. Figure 1b reveals that there are some discrepancies between the estimated discrimination parameters $\hat{\boldsymbol{a}}$ under the LHL-3PL and the LHL-2PL model, whereas the estimated difficulty parameters $\hat{\boldsymbol{b}}$ are very close (Fig. 1c). The sample correlations between the estimates under the two models are 0.177 and 0.969 for $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$ respectively (Table 3).

The sample correlation matrix of $\hat{\boldsymbol{a}}$, $\hat{\boldsymbol{b}}$, $\hat{\boldsymbol{c}}$ and $\boldsymbol{\pi}$ under LHL-2PL and LHL-3PL given in Table 4 shows that $\boldsymbol{\pi}$ is highly correlated with $\hat{\boldsymbol{b}}$, and is negatively correlated (although the correlation is moderate) with $\hat{\boldsymbol{a}}$ under LHL-3PL while almost uncorrelated under LHL-2PL. For $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$, there is a moderate correlation

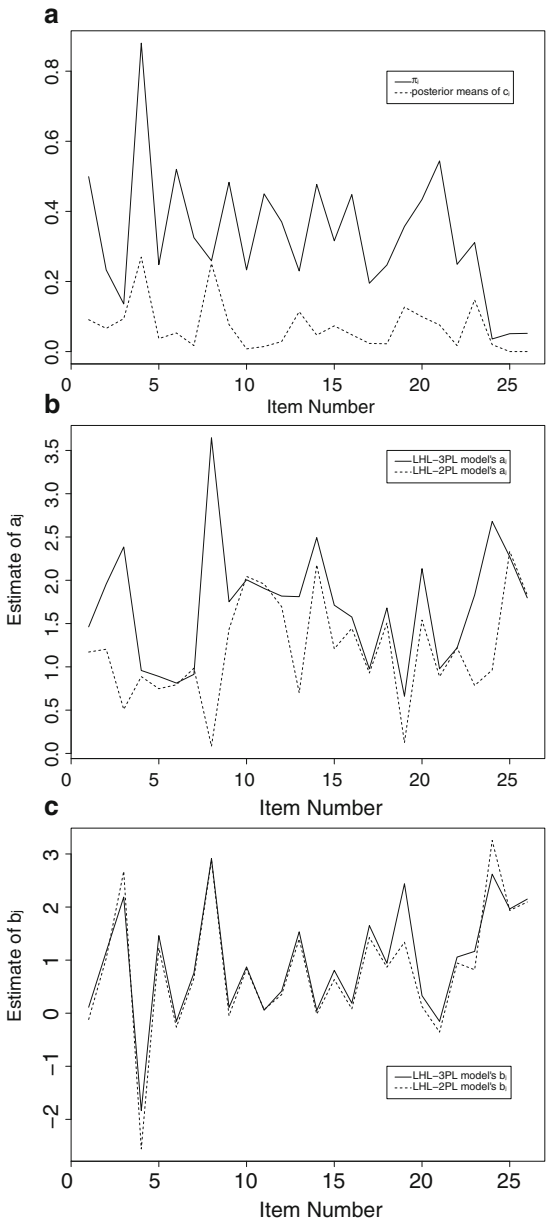


Fig. 1 (**a**) Plots of $\boldsymbol{\pi}$ and $\hat{\boldsymbol{c}}$, for $j = 1, \ldots, 26$; (**b**) plots of $\hat{\boldsymbol{a}}$ under LHL-3PL and LHL-2PL; (**c**) plots of $\hat{\boldsymbol{b}}$ under LHL-3PL and LHL-2PL

under LHL-3PL, whereas there is a low and negative correlation under LHL-2PL. For $\boldsymbol{\pi}$ and $\hat{\boldsymbol{c}}$, they are moderately correlated.

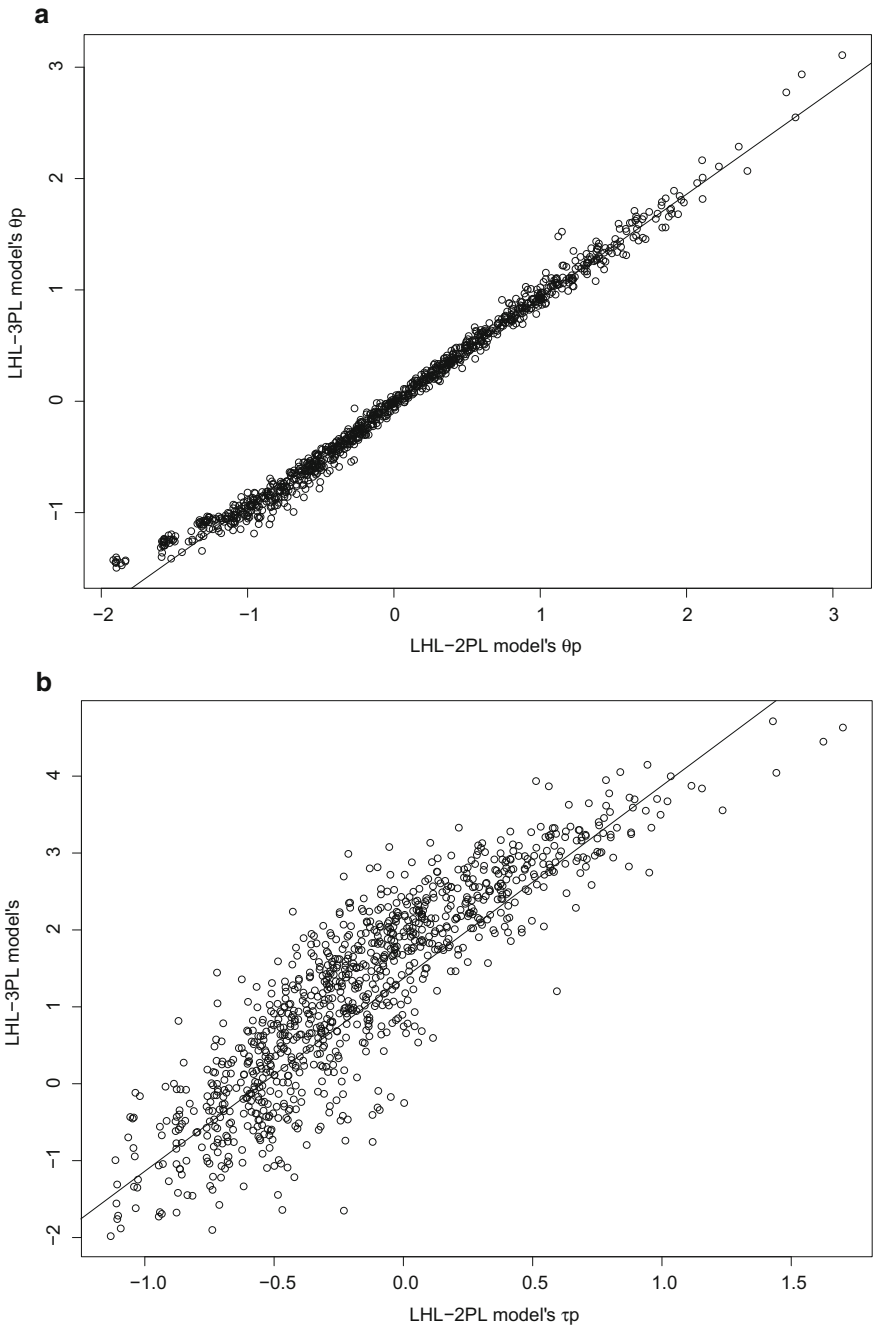


Fig. 2 (**a**) Scatter plot of $\hat{\boldsymbol{\theta}}$ under LHL-3PL against LHL-2PL; (**b**) Scatter plot of $\hat{\boldsymbol{\tau}}$ under LHL-3PL against LHL-2PL

Table 3 Sample correlations between the estimates under LHL-3PL and LHL-2PL

	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\tau}}$	$\hat{\boldsymbol{a}}$	$\hat{\boldsymbol{b}}$
Correlation	0.994	0.882	0.177	0.969

Table 4 Sample correlations of the estimates for LHL-3PL, with their counterparts for LHL-2PL enclosed by parentheses

	$\hat{\boldsymbol{a}}$	$\hat{\boldsymbol{b}}$	$\hat{\boldsymbol{c}}$
$\boldsymbol{\pi}$	−0.387(−0.067)	−0.877(−0.919)	0.492
$\hat{\boldsymbol{a}}$		0.417(−0.204)	0.150
$\hat{\boldsymbol{b}}$			−0.132

Figure 2a shows that the estimated $\hat{\boldsymbol{\theta}}$ under both models yields very similar results. Figure 2b, however, shows that there is a larger difference between the estimated examinee-specific threshold parameters. Indeed, the variations of $\hat{\boldsymbol{\tau}}$ under LHL-3PL are much larger than those of LHL-2PL. This may be due to the inclusion of the extra pseudo-guessing parameters in the LHL-3PL model. The sample correlations between the estimates under LHL-3PL and LHL-2PL are 0.994 and 0.882 for $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\tau}}$ respectively (Table 3).

Figure 1a and b reveals that item 8 has a π that is very close to its c-parameter estimate and it has very different a-parameter estimates in the LHL-2PL and LHL-3PL. We therefore compute the estimated probability that item 8 is answered correctly in these four models. We first consider the LHL-3PL model. This is done as follows. Recall that we produce 40,000 MCMC draws with the first 10,000 draws as burn-in. For $p = 1, \ldots, P$, for each draw after burn-in, we compute the probability that $\{Y_{p8} = 1\}$ using Eq. (2), then we take the average over all the last $30,000$ draws to get an estimate of the probability that $\{Y_{p8} = 1\}$. Then, we take the average over $p = 1, \ldots, P$, to get the estimate of the probability that item 8 is answered correctly. We repeat the computation for the other 3 models. The estimated values are 0.26642, 0.25783, 0.26285, and 0.25860 in the 2PL, 3PL, LHL-2PL, and LHL-3PL models, respectively. Since $\pi_8 = 0.259$, we see that the estimate in the LHL-3PL model is closest to π_8. However, the interpretations under the LHL-2LP and LHL-3PL models are quite different. In the LHL-3PL model, item 8 has the highest b-parameter estimate, meaning that it is the most difficult one, and most examinees answer it correctly just by guessing. This may or may not be true, and deserves a further study by putting some stronger priors on the $\boldsymbol{c}$-parameter instead of using a two-layer hierarchical prior in this study to reduce the impact of the prior settings.

5 Concluding Remarks

In this study, we extend the LHL-2PL model to the LHL-3PL model by adding a pseudo-guessing parameter. Then, we apply the LHL-3PL model to the physics examination data of the Department Required Test for college entrance in Taiwan. The test consists of three types of questions, including multiple-choice, multiple-

response, and calculation problems. The percent of examinees who responded correctly are the lowest for the two calculation problems. The estimated pseudo-guessing parameters for the multiple-choice and multiple-response questions range from 0.0077 to 0.2694, indicating some evidence of random guessing. This may due to the fact that examinees often know enough about the subject to eliminate some of the incorrect choices. Therefore, guessing from among the remaining options is likely to help them beat the odds of random guessing. We found that the estimated ability parameters almost unaffected by adding a pseudo-guessing parameter to the model. The changes in the estimated difficulty parameters are also slim. Changes are mainly in some of the estimated discrimination parameters and many of the estimated examinee-specific threshold parameters for the speededness effect. In sum, we find some evidence for the LHL mechanism as well as for random guessing.

In the LHL-3PL model, we consider the case that all the examinees share the same speededness rate λ. It is interesting to relax the assumption in a further study. Another interesting future work is to put some stronger priors on the $\boldsymbol{c}$-parameter.

Acknowledgements The research was supported by Academia Sinica and the Ministry of Science and Technology of the Republic of China under grant number MOST 102-2118-M-001 -007 -MY2. The authors would like to thank the co-editor, Professor Wen-Chung Wang, and Dr. Yu-Wei Chang for their helpful comments and suggestions, and the College Entrance Examination Center (CEEC) for providing the data.

References

Agresti, A. (2002). *Categorical data analysis*. Hoboken, NJ: Wiley.

Angoff, W. H. (1989). Does guessing really help? *Journal of Educational Measurement, 26*, 323–336.

Bejar, I. I. (1985). *Test speededness under number-right scoring: An analysis of the test of English as a foreign language* (Research Rep. RR-85-11). Princeton: Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

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*, 331–348.

Boughton, K. A., & Yamamoto, K. (2007). A HYBRID model for test speededness. In M. von Davier, & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 147–156). New York: Springer.

Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika, 73*, 209–230.

Chang, Y.-W., Tsai, R.-C., & Hsu, N.-J. (2014). A speeded item response model: Leave the harder till later. *Psychometrika, 79*, 255–274.

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

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

Fox, J.-P. (2010). *Bayesian item response modeling-theory and applications*. New York: Springer.

Goegebeur, Y., De Boeck, P., Wollack, J. A., & Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika, 73*, 65–87.

Huang, H.-Y., & Hung, S.-P. (2010). Implementation and application of Bayesian three-level IRT random intercept latent regression model. *Chinese Journal of Psychology, 52*, 309–326. (in Chinese)

Lee, Y.-H., & Ying, Z. (2015). A mixture cure-rate model for responses and response times in time-limit tests. *Psychometrika, 80*, 748–775.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*, 3–21.

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement, 12*, 7–11.

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement, 34*, 521–538.

Mislevy, R. L. (1986). Bayes modal estimation in item response theory. *Psychometrika, 51*, 177–195.

Rost, J. (1990). Rasch models in latent classes: an integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271–282.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*, 298–321.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B, 64*, 583–616.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7*, 175–192.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the twoparameter logistic model. *Psychometrika, 50*, 349–364.

Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51*, 589–601.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287–308.

Wang, W.-C. (2004). Rasch measurement theory and application in education and psychology. *Journal of Education and Psychology, 27*, 637–694 (in Chinese).

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

Yamamoto, K. (1989). *HYBRID model of IRT and latent class models* (ETS Research Rep. No. RR-89-41). Princeton: Educational Testing Service.

Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYBRID model* (TOEFL Technical Rep. No. TR-10). Princeton: Educational Testing Service.

Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the hybrid model. In J. Rost (Ed.), *Applications of latent trait and latent class models in the social sciences* (pp. 89–98). Munster: Waxmann.

An Application of a Random Mixture Nominal Item Response Model for Investigating Instruction Effects

Hye-Jeong Choi, Allan S. Cohen, and Brian A. Bottge

Abstract The purpose of this study was to apply a random item mixture nominal item response model (RIM-MixNRM) for investigating instruction effects. The host study design was a pre-test-and-post-test, school-based cluster randomized trial. A RIM-MixNRM was used to identify students' error patterns in mathematics at the pre-test and the post-test. Instruction effects were investigated in terms of students' transitioning in error patterns. That is, we compared students' error patterns in the Enhanced Anchored Instruction (EAI) condition with students' error patterns in a business-as-usual (BAU) instructional condition following each instruction. We also compared error patterns of students with math disabilities and students without math disabilities following the two types of instruction.

Keywords Random item model • Mixture IRT model • Nominal responses model

1 Introduction

Mixture item response theory (MixIRT) models have been used for modeling population heterogeneity (e.g., Mislevy & Verhelst 1990; Rost 1990). Most mixture models consider only persons random but items fixed. Random item IRT models (De Boeck 2008), however, consider both item and person parameters as random. This is more appealing as (1) both items and persons are typically assumed to be random samples from some population and (2) treating both items and persons as random permits inclusion of covariates on both item and person parameters to help explain differences in both item and examinee parameters (Van den Noortgate, De Boeck, & Meulders, 2003; Wang 2011).

Also, to date, most MixIRT models have primarily focused on dichotomously or polytomously scored items. MixIRT models can be usefully applied to nominal

H.-J. Choi (✉) • A.S. Cohen
University of Georgia, Aderhold Hall 125, Athens, GA 30601, USA
e-mail: hjchoi1@uga.edu

B.A. Bottge
University of Kentucky, 222 Taylor Education Building, Lexington, KY, USA

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer
Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_4

responses items. For example, MixIRT models can effectively model to capture information about specific error patterns which individual distractors for multiple choice items may contain.

The purpose of this study was to apply a random item mixture nominal response model to an empirical data set for investigating instructional effects. An important benefit of such a model is that it is possible to explicitly model randomness of item and ability parameters as well as specific aspects of students' response patterns. We provide a brief description of the random item mixture nominal model (RIM-MixNRM) and a simulation study to evaluate the quality of the estimation method. Then, we provide an empirical example in which a RIM-MixNRM is applied to mathematic test data to investigate effects of an experimental instruction on students' error patterns on fractions computation.

2 A Random Item Mixture Nominal Response Model

The probability of selecting individual categories in an item with two or more nominal categories can be written as a linear function of item category and person parameters. Bock (1972), for instance, introduces a nominal model in which the probability of selecting category k of item i, $P_{ik}(\theta_j)$, is defined as a multinomial logistic function:

$$P_{ik}(\theta_j) = \frac{\exp(\lambda_{ik}\theta_j + \zeta_{ik})}{\sum_{k=1}^{K} \exp(\lambda_{ik}\theta_j + \zeta_{ik})}, \tag{1}$$

where

$i = 1, \ldots, n$ items,
$k = 1, \ldots, K$ response categories,
$j = 1, \ldots, N$ examinees,
ζ_{ik} denotes the intercept for category k of item i,
λ_{ik} denotes the slope for category k of item i, and
θ_j denotes the person parameter of person j.

Bolt, Cohen, and Wollack (2001) extended this model to a mixture nominal IRT model as a way of detecting heterogeneity in a population. In doing so, Bolt et al. (2001) included a class-specific category intercept parameter to specify the propensity of selecting a given category of item i for members of latent class g. The class-specific probability of a response is given by

$$P_{gik}(\theta_j) = \frac{\exp(\lambda_{ik}\theta_j + \zeta_{gik})}{\sum_{k=1}^{K} \exp(\lambda_{ik}\theta_j + \zeta_{gik})}, \tag{2}$$

with marginal probability

$$P_{ik}(\theta_j) = \sum_{g}^{G} \pi_g P_{gik}(\theta_j) \,, \tag{3}$$

where ζ_{gik} denotes the class-specific category intercept, $g = 1 \ldots, \mathrm{G}$ latent classes, and π_g mixing proportion ($\sum_g^G \pi_g = 1$). For resolving an indeterminacy for the item category parameters, constraints of $\sum_k^K \lambda_{ik} = 0$ and $\sum_k^K \zeta_{gik} = 0$ were set for all items and all classes.

Bolt et al. (2001) applied Markov chain Monte Carlo (MCMC) algorithm to estimate the model in a general hierarchical framework and a fully Bayesian approach as implemented in the computer software computer software WinBUGS. They used following conjugate priors:

$$
\begin{aligned}
& c_j \sim \mathit{Multinomial}(1|\pi_1, \ldots, \pi_G) \\
& \pi = (\pi_1, \ldots, \pi_G) \sim \mathit{Dirichlet}(\alpha_1, \ldots, \alpha_G) \\
& \theta_j|c_j = g \sim N(\mu_{\theta_g}, \sigma^2_{\theta_g}) \\
& \lambda_{ik} \sim N(\mu_\lambda, \sigma^2_\lambda) \\
& \zeta_{gik} \sim N(\mu_{\zeta_g}, \sigma^2_{\zeta_g}).
\end{aligned}
$$

In their model, however, item parameters were treated as fixed as in the conventional mixture item response models. In the current study, we extended their model to a model where both item and person parameters are treated as random.

3 Simulation Study

The simulation study described below was designed to examine the behavior of the RIM-MixNRM under practical testing conditions.

3.1 Simulation Design

Hundred sets of 20 four-choice item responses were simulated from a standard normal distributions, $N(0, 1)$. Six hundred examinees for three latent classes were simulated and mixing proportions were 0.33 and ability was generated as $N(0, 1)$ for each class. Item parameter estimates adopted from Bolt et al. (2001) were used to select item generating parameters. Generating values for model parameters are given in Table 1. As can be seen in Eq. (2), in this particular RIM-MixNRM, ζ_{gik} is the parameter to distinguish latent classes.

The parameters for hyperpriors were used: $\alpha_1 = \ldots = \alpha_G = 0.5$; $\mu_{\theta_g} \sim N(0, 1)$; $1/\sigma^2_{\theta_g} \sim \mathit{Gamma}(2, 4)$; $\mu_\lambda \sim N(0, 1)$; $1/\sigma^2_\lambda \sim \mathit{Gamma}(2, 4)$, $\mu_{\zeta_g} \sim N(0, 1)$; $1/\sigma^2_{\zeta_g} \sim \mathit{Gamma}(2, 4)$. These parameters were similar with ones used by Bolt et al. (2001) and only provided minimum information for each parameter. In addition to

Table 1 Item parameter for generating data sets for simulation study

	Slope				Threshold											
					Class 1				Class 2				Class 3			
	λ_1	λ_2	λ_3	λ_4	ζ_1	ζ_2	ζ_3	ζ_4	ζ_1	ζ_2	ζ_3	ζ_4	ζ_1	ζ_2	ζ_3	ζ_4
1	1.01	0.30	−0.14	−1.17	0.74	0.20	0.04	−0.98	0.74	2.70	0.04	−0.98	0.74	0.20	0.04	−0.98
2	0.89	−0.21	0.39	−1.07	0.44	0.30	−0.84	0.10	0.44	2.80	−0.84	0.10	0.44	0.30	−0.84	0.10
3	1.69	−0.29	−0.51	−0.89	−0.57	0.34	0.33	−0.10	−0.57	2.84	0.33	−0.10	−0.57	0.34	0.33	−0.10
4	0.99	−0.32	0.13	−0.80	−0.22	0.78	0.31	−0.87	−0.22	3.28	0.31	−0.87	−0.22	0.78	0.31	−0.87
5	0.73	−0.42	0.31	−0.62	−0.97	0.50	−0.93	1.40	−0.97	3.00	−0.93	1.40	−0.97	0.50	−0.93	1.40
6	1.62	−0.53	−0.54	−0.55	−1.86	0.02	2.51	−0.67	−1.86	2.52	2.51	−0.67	−1.86	0.02	2.51	−0.67
7	0.95	0.29	0.38	−1.62	−1.11	0.20	2.69	−1.78	−1.11	0.20	2.69	−1.78	−1.11	0.20	5.19	−1.78
8	1.06	−0.61	0.50	−0.95	0.42	−0.69	1.41	−1.14	0.42	−0.69	1.41	−1.14	0.42	−0.69	3.91	−1.14
9	1.20	−0.39	0.27	−1.08	1.25	−1.02	1.34	−1.57	1.25	−1.02	1.34	−1.57	1.25	−1.02	3.84	−1.57
10	0.91	0.12	0.78	−1.81	−0.77	1.36	−0.35	−0.24	−0.77	1.36	−0.35	−0.24	−0.77	1.36	2.15	−0.24
11	0.91	0.46	−0.43	−0.94	1.25	−0.16	−0.58	−0.51	1.25	−0.16	−0.58	−0.51	1.25	−0.16	1.92	−0.51
12	1.42	0.03	0.34	−1.79	0.16	0.35	0.22	−0.73	0.16	0.35	0.22	−0.73	0.16	0.35	2.72	−0.73
13	1.09	−0.23	−0.32	−0.54	1.63	−0.84	−0.12	−0.67	1.63	−0.84	−0.12	−0.67	1.63	−0.84	2.38	−0.67
14	1.19	−0.24	0.36	−1.31	−2.20	0.64	0.72	0.84	−2.20	0.64	0.72	0.84	−2.20	0.64	3.22	0.84
15	0.90	0.46	−0.40	−0.96	0.45	−0.54	0.19	−0.10	0.45	−0.54	0.19	−0.10	0.45	−0.54	2.69	−0.10
16	0.93	−0.52	0.32	−0.73	0.06	0.54	0.12	−0.72	0.06	0.54	0.12	−0.72	0.06	0.54	0.12	−0.72
17	1.34	−0.29	−0.11	−0.94	0.00	0.14	1.09	−1.23	0.00	0.14	1.09	−1.23	0.00	0.14	1.09	−1.23
18	1.64	0.21	−0.84	−1.01	−0.67	−0.21	1.42	−0.54	−0.67	−0.21	1.42	−0.54	−0.67	−0.21	1.42	−0.54
19	1.30	0.12	−0.62	−0.80	−0.78	−0.32	0.48	0.62	−0.78	−0.32	0.48	0.62	−0.78	−0.32	0.48	0.62
20	1.06	−0.29	−0.35	−0.42	−0.83	0.12	1.23	−0.52	−0.83	0.12	1.23	−0.52	−0.83	0.12	1.23	−0.52

$\sum_k^K \lambda_{ik} = 0$, $\sum_k^K \zeta_{gik} = 0$, and $\sum_g^G \pi_g = 1$, for identification, μ_θ and σ_θ set to zero and one for the first class. These priors, hyperpriors, and constraints were also used for analyzing the empirical data set in the later section. The computer software computer software WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2007) was used for both simulation and empirical studies.

Convergence of the MCMC algorithm was examined using the Geweke test (1992) with a single chain as implemented in the computer program Convergence Diagnosis and Output Analysis for MCMC (CODA: Plummer, Best, Cowles, & Vines, 2006). Based on the Geweke test (1992) with a single chain and plots of autocorrelations, kernel density estimates of the marginal posterior distributions, and history plots of draws from posteriors, a burn-in of 3000 iterations was sufficient to achieve stationarity for all parameter estimates. Following this burn-in, an additional 5000 iterations were drawn to obtain estimates for each of the posterior distributions of model parameter estimates.

3.2 Simulation Study Results

3.2.0.1 Model Selection To investigate whether model fit indices could identify the correct number of latent classes, one- to four-class RIM-MixNRMs were fit to the data sets. The Bayesian information criterion (BIC: Schwarz 1978) and Akaike's information criterion (AIC: Akaike 1974) were chosen as model fit indices because of their popularity among researchers. Both indices were able to identify the correct model, that is, both indicated the three-class model was the best fit model for 93 of 100 replications. For the remaining replications, BIC suggested a two-class model and AIC a four-class model.

Label Switching Label switching is a well-known problem in finite mixture modeling. Two types of label switching can occur with mixture modeling (Cho, Cohen, & Kim, 2013). The first occurs over a single MCMC chain: the labels of the latent classes switch during estimation. The second type of label switching may be observed when labels switch between multiple data sets or multiple analyses in both Bayesian and maximum likelihood estimation. In the context of a simulation study, one needs to be aware of the possibility of label switching, as when labels switch on different replicate data sets, this may cause confusion when interpreting results. In the current study, the possibility of occurrence of label switching was investigated by inspecting profiles of item estimates across latent classes. When label switching was detected, latent classes were renamed by matching the profiles of parameter estimates across replications before calculating bias, mean squared error (MSE), and classification accuracy rates.

Recovery of Parameters A recovery analysis was done to determine whether the MCMC algorithm accurately recovered the model parameters of the RIM-MixNRM. In addition to inspection of label switching, parameter estimates had been placed on the metric of the generating parameters and then bias and MSEs

of parameter estimates were calculated. Correlations between generation values and estimates also used for the recovery analysis.

Results showed that most of the parameters of the RIM-MixNRM were recovered well: bias of item slopes, person ability and mixing proportion were about or less than 0.05; MSE were about or less than 0.16; and correlations were about or higher than 0.93. Correlations between item threshold parameters and estimates, ζ, was 0.94, but bias and MSE were -0.16 and 0.16, respectively, and appeared to depart slightly from generated values than other parameters. The RIM-MixNRM correctly classified examinees into their true (i.e., simulated) classes 87.85 % of the time.

4 Empirical Study: Instruction Effects on Students' Fractions Computation

In this section, we illustrate how a RIM-MixNRM can be used to help investigate effects of instruction on students' learning process. In this example, students' error patterns were examined on a test of fractions computation in a multi-year cluster randomized instructional intervention. The main purpose of the study was to investigate an experimental instructional condition effects on students' error patterns in computing fractions.

4.1 Data Description

Study Design The host study design was a school-based cluster randomized trial. Participants included 446 middle school students in Grades 6–8 in 25 general education math classrooms in 12 middle schools in and around a large metropolitan area in the Southeast. Students were randomly assigned to an experimental instructional condition (N = 214) or to a business-as-usual (BAU) condition (N = 232). There were 123 students with learning disabilities and 323 students without learning disabilities in the study.

The experimental condition implemented Enhanced Anchored Instruction (EAI; Bottge, Ma, Toland, Gassaway, & Butler, 2012). EAI was designed for use in helping to improve computation and problem solving skills of adolescents, including low-performing students with learning disabilities by including practical, hands-on applications to help students visualize the abstract concepts present in the problem. Teachers ask probing questions and offer instructional guidance to students as they view the video and help them identify relevant information to solve the problem. This eliminates the need for reading, a skill many low-achieving math students also lack.

Fractions Computation Test A Fractions Computation Test (FCT) consisting of 20 partial credit items (14-addition and 6-subtraction items) was designed by Bottge

et al. (2012) to measure students' ability to add and subtract simple fractions and mixed numbers with like and unlike denominators. The FCT was administered for the pre-test and the post-test. Math education experts identified 11 types of errors from students' incorrect responses to these items. The most common were *Combining* (*C*) and *Selecting Denominator* (*SD*). The remaining nine other types of errors occurred less frequently and were combined into a single *Other* (*O*) for this study. In the current study, the focus was on these three types of errors as they reflect students' misunderstandings about computing with fractions as well as the correct response (i.e., *No* errors). *Combining* and *SelectingDenominator* errors are described below:

- Combining (C): Student combines numerators and combines denominators, consistently applying the same operation to numerator and denominator.
- Select Denominator (SD): Student selects one of the denominators listed in the problem and makes no attempt to make equivalent fraction. Denominator given in the answer must be present in the problem.

4.2 Model Estimation

To investigate EAI effects on students' error patterns in fractions computation, we applied RIM-MixNRMs to take into account randomness in students and items parameters. The instructional method (i.e., EAI vs BAU) and students' math learning disability status (MD) were included in the model as covariates to predict the latent class membership as this could reflect of EAI effects on students' error patterns. This was done by substituting π_g in Eq. (3) with $\pi_{g|\mathbf{x}}$ as given by

$$\pi_{g|\mathbf{x}} = \frac{\exp(\beta_{g0} + \boldsymbol{\beta}_g' X)}{\sum_{g=1}^{G} \exp(\beta_{g0} + \boldsymbol{\beta}_g' X)} \tag{4}$$

where β_{g0} denotes an intercept for class g, and $\boldsymbol{\beta}_g = (\beta_{g1}, \beta_{g2}, \ldots, \beta_{gp})$ is a vector of logistic regression coefficients of covariates in the model. For this study, those covariates were the instructional method and students' math learning disability.

For those $\beta's$, normal distributions with mean of zero and standard deviation of 10 were used as conjugate priors and for rest parameters, the same priors were used as in the simulation study. For identification, the first class was used as a reference group.

The Geweke test, for convergence diagnosis, indicated that a burn-in of 4000 iterations was sufficient to achieve stationarity for all parameters. A subsequent 6000 iterations were used for estimating the model parameters. An exploratory analysis was applied to determine the number of latent classes in the data. That is, RIM-MixNRMs with from one- to five-class were fit to the pre- and post-test data. Based on BIC results, a four-class and a three-class RIM-MixNRMs for pre-test and post-test data, respectively were chosen for this study.

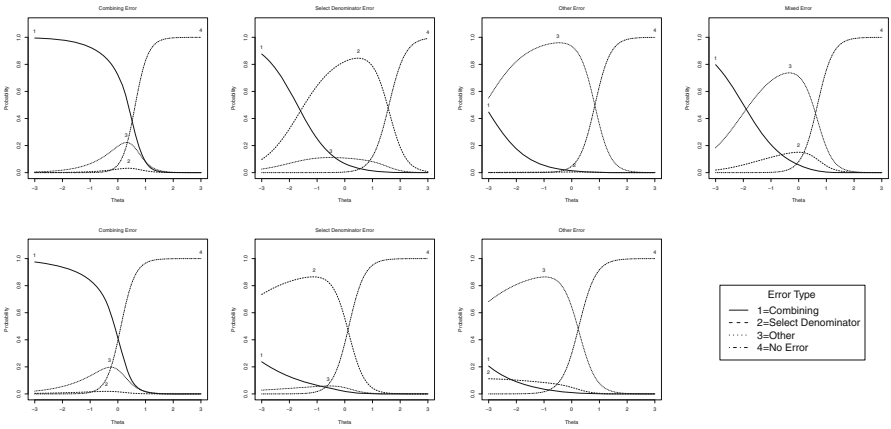


Fig. 1 Item category characteristic curves for item 16 showing latent class differences in students' error patterns on fractions computation

4.3 Results

4.3.1 Characteristics of Latent Classes

Item category characteristic curves (ICCC's) for Item 16 are shown in Fig. 1 for each latent class. These plots illustrate differences in types of errors made by students of the individual latent classes. The plots in the upper panel are for the pre-test and the plots in the lower panel are for the post-test. Students in all classes had a greater probability of not making any errors as they possessed more ability (i.e., Category 4). However, there were distinct error patterns which students in middle or lower ability in individual latent classes tended to make. Some students in middle or lower ability tended to mistakenly combine each numerator and each denominator (i.e., Category 1), some had a greater probability of making an error of selecting denominator (i.e., Category 2), some had a greater probability of making other errors (i.e., Category 3), and others had a greater probability of making either combining or selecting denominators (i.e., Category 1 or 2). Based on these patterns, each class is labeled as *Combining*, *SD*, *Other*, or *Mixed* shown in Fig. 1. These distinct differences can be interpreted as reflecting students' error pattern on the FCT.

4.3.2 Instruction Effects

Table 2 presents a cross-tabulation of the frequencies of students in each latent class on the pre- and the post-test. This shows a general pattern of students' transitioning in class membership from the pre-test to the post-test. To investigate effects of students' learning disability status and instructional type on such transitioning, those were included as covariates. On the pre-test, neither types of instruc-

Table 2 Transitioning pre-test to post-test latent classes

	Post-test			
Pre-test	Combine	SD	Other	Total
Combine	57	27	82	166
SD	3	25	33	61
Other	12	1	96	109
Mixed	19	16	75	110
Total	91	69	286	446

tion nor students' learning disability status did significantly impact on students membership in latent classes except that students with learning disability had significantly lower odds of belonging to *Other* error class than *Combining* error class (i.e., $\beta_{20} = -1.11, 0.33$ times). After the intervention, however, EAI had significant impact on students' error patterns on fraction computation. Students in EAI had significantly higher odds of belonging to *Other* or *SD* error classes than *Combining* error class (i.e., $\beta_{22} = 1.06, 2.89$ times and $\beta_{32} = 0.85, 2.34$ times, respectively). After the instruction, students might better understand about denominators and could distinguish them from numerators but still not fully understand the concept of common denominator in fractions.

5 Conclusion and Discussions

It is not uncommon that researchers or practitioners design an instrument to require nominal responses with a specific purpose in educational and psychological research area. For instance, in creating multiple choice items, item writers typically construct distractors in order to represent specific errors students might make. Nominal IRT models can be used to obtain information regarding these errors. Further, a mixture nominal IRT model can be used to take into account population heterogeneity; however, it does not consider randomness in items. In this study, we used a RIM-MixNRM in which both items and person parameters were considered a random sample from a population and taken into account their randomness. Results from a simulation study suggested that the model parameters were well recovered and both AIC and BIC provided useful information for model selection. Results from the middle school fractions computation data revealed there were four latent classes and three latent classes on the pre-test and the post-test, respectively, which could reflect students' error pattern on fractions computation. The results also show that instructional type had an significant impact on transitioning these error patterns subsequent to an instructional intervention. It is also possible to include item covariates. Inclusion of a Q-matrix (e.g., Tatsuoka 1983), for instance, as a covariate for individual categories of an item could be implemented to describe components of knowledge required for correctly answering a given question.

Acknowledgements The data used in the article were collected with the following support: the U.S. Department of Education, Institute of Education Sciences, PR Number H324A090179.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

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

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture model for multiple choice data. *Journal of Educational and Behavioral Statistics, 26*, 381–409.

Bottge, B., Ma, X., Toland, M., Gassaway, L., & Butler, M. (2012). *Effects of Enhanced Anchored Instruction on middle school students with disabilities in math*, Department of Early Childhood, Special Education, and Rehabilitation Counseling, University of Kentucky, Lexington, KY.

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

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*, 533–559.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting* (pp. 169–194). Oxford: Oxford University Press.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195–215.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News, 6*(1), 7–11.

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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2007). WinBUGS, 1.4 [Computer program].

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345–354.

Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics, 28*(4), 369–386.

Wang, A. (2011). *A mixture cross-classification IRT model for test speededness*. Unpublished doctoral dissertation, University of Georgia.

Item Response Theory Models for Multidimensional Ranking Items

Wen-Chung Wang, Xuelan Qiu, Chia-Wen Chen, and Sage Ro

Abstract Multidimensional ranking items, in which different statements aim at different latent traits, are commonly used to measure noncognitive latent traits (e.g., career interests, attitudes, and personality). In this study, we developed two new item response theory models for multidimensional ranking items that yield statement utilities and person measures. Simulations were conducted to evaluate the parameter recovery of the two new models, and the results indicated that the parameters were recovered well by using the freeware Just Another Gibson Sampler (JAGS). An empirical example of behaviors in workplaces was provided.

Keywords Item response theory • Ranking items • Pairwise comparison • Rasch measurement • Ipsative tests

The use of ranking items as an instrument has a long history in the marketing, economics, and politics literature. A sample item retrieved is the following: Please rank four psychology career areas according to your career preferences: Academic, Clinical, Educational, and Industrial (Maydeu-Olivares & Böckenholt 2005). There are two major approaches to modeling ranking items. One is the binary-coding approach in which a ranking pattern is recoded as a series of paired comparisons, and then models, such as Thurstone's (1927) comparative judgment models (Thurstone 1931) or the Bradley–Terry–Luce random utility model (Bradley & Terry 1952; Luce 1959), are applied to the paired comparisons. For example, a ranking item with three statements, r_1, r_2, and r_3, can be recoded into three pairwise-comparison items, each with two statements: (r_1, r_2), (r_2, r_3), and (r_1, r_3). The probability of intransitive pairwise-comparison patterns should be set at zero to maintain the rank orders (e.g., Brown & Maydeu-Olivares 2011; Maydeu-Olivares & Böckenholt 2005; Maydeu-Olivares & Brown 2010). This approach

W.-C. Wang (✉) • X. Qiu • C.-W. Chen
Department of Psychological Studies, The Education University of Hong Kong,
10 Lo Ping Road, Tai Po, New Territories, Hong Kong
e-mail: wcwang@ied.edu.hk

S. Ro
IBM, 650 Third Avenue South, Minneapolis, MN, 55402 USA

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer
Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_5

becomes cumbersome when there are more than four statements in a ranking item. The other approach directly describes the probability of each ranking pattern without performing binary coding, such as the exploded logit model (Allison & Christakis 1994; Chapman & Staelin 1982; Punj & Staelin 1978; also called the rank-ordered logit model; Beggs, Cardell & Hausman 1981; Hausman & Ruud 1987) that has been widely used in marketing and economics research.

In these two approaches to ranking data, all statements involved are assumed to measure the same latent trait, which makes person measures invisible (not measurable). This type of ranking item is referred to as a unidimensional ranking item. Ranking items can consist of statements that measure different latent traits, which are referred to as multidimensional ranking items. For instance, in the Study of Values Part II (Kopelman, Rovenpor, & Guan 2003), respondents were asked to rank four statements according to their views of da Vinci's masterpiece "The Last Supper": (a) spiritual aspirations and emotions, (b) priceless and irreplaceable, (c) da Vinci's versatility, and (d) quintessence of harmony and design. These four statements measure four latent traits, respectively, religious value, economic value, theoretical value, and aesthetic value. Other psychological inventories that employ multidimensional ranking items include the Occupational Preference Questionnaire (Saville, Sik, Nyfield, Hackston & MacIver 1996), the Occupational Personality Questionnaire (SHL 2006), and others (Salgado & Tauriz 2014).

A fundamental feature of scoring multidimensional ranking items is the indeterminacy of the scale origin. Although a person may find all statements in a ranking item attractive and another person may find them unattractive, the individuals' comparative judgments can be identical, and thus, the individuals' absolute assessments of the statements cannot be inferred (Böckenholt 2004). That is, ranking items yield ipsative measures (Cattell 1944), and therefore, a comparison between individuals at "absolute" levels of latent traits is impossible (de Vries & van der Ark 2008; Meade 2004).

There have been attempts to develop item response theory (IRT) models for multidimensional ranking items. Brown and Maydeu-Olivares (2011, 2012, 2013) applied the binary coding approach within the IRT framework and declared that their model resolves the fundamental problem of the indeterminacy of the scale origin and suggests absolute assessments of the latent traits. Actually, the model does not possess measurement properties (e.g., monotone homogeneity and double monotonicity; Mokken 1971) and cannot yield meaningful measures for any comparison (Chen & Wang 2013a). Acknowledging the ipsative nature (indeterminacy of scale origin) in ranking items, Chen and Wang developed a Rasch ipsative model (RIM) for multidimensional ranking items with two statements in each ranking item (i.e., pairwise-comparison items). As a member of the family of Rasch models, the RIM inherits the good measurement properties of specific objectivity and sufficient statistics for item and person parameters.

The RIM is limited to multidimensional pairwise-comparison items. In this study, we extended the RIM and proposed two IRT approaches to multidimensional ranking items (consisting of more than two statements in each item). The first approach was inspired by the classical exploded logit model, which formulates the

probability of a ranking pattern as the product of the first-choice probabilities for successive remaining alternatives. The other approach formulates the probability of a ranking pattern as the product of a series of pairwise-comparison probabilities. Both approaches, although built on different foundations, include the RIM as a special case when each ranking item has only two statements. The remainder of this paper is organized as follows. First, the RIM is briefly introduced. Second, the two approaches to multidimensional ranking items are illustrated. Third, the results of the simulation studies that were conducted to evaluate the parameter recovery of the new models are outlined. Fourth, the new models are applied to empirical data to demonstrate the models' implications and applications. Finally, conclusions about the new models are drawn, and directions for future studies are discussed.

1 The Rasch Ipsative Model for Multidimensional Pairwise-Comparison Items

Let there be J statements $r_1, r_2, \ldots, r_J$ in a test with utility values of $u_{r_1}, u_{r_2}, \ldots u_{r_J}$, respectively. Let η_{dn} denote person n's "absolute" level on latent trait d ($d = 1, \ldots, D$). Usually, each latent trait is measured by multiple statements; thus, J is much larger than D. Due to the ipsative nature of pairwise-comparison items, it is impossible to identify η_{dn}. Let $\overline{\eta}_n$ be the mean across D latent traits for person n (i.e., $\overline{\eta}_n = \sum_{d=1}^{D} \eta_{dn}/D$). Then, for each person n, we have the following:

$$\theta_{dn} = \eta_{dn} - \overline{\eta}_n, \tag{1}$$

where the θ variables sum to zero across D latent traits for every person (i.e., $\sum_{d=1}^{D} \theta_{dn} = 0$). When two statements, r_j and $r_{j'}$, are compared in a pairwise-comparison item, the log-odds of selecting r_j over $r_{j'}$ for person n according to the RIM (Chen & Wang 2013a) are defined as:

$$\log\left(\frac{P_{rj}\left|\left(r_j, r_{j'}\right), n\right.}{P_{rj'}\left|\left(r_j, r_{j'}\right), n\right.}\right) = \left(\theta_{dn} + u_{r_j}\right) - \left(\theta_{d'n} + u_{r_{j'}}\right), \tag{2}$$

where $P_{rj}\left|\left(r_j, r_{j'}\right), n\right.$ and $P_{rj'}\left|\left(r_j, r_{j'}\right), n\right.$ are the probabilities of selecting r_j and $r_{j'}$ for person n, respectively, θ_{dn} and $\theta_{d'n}$ are the "relative" levels of latent traits d and d' for person n that are measured by statements r_j and $r_{j'}$, respectively, and u_{r_j} and $u_{r_{j'}}$ are the utilities for r_j and $r_{j'}$, respectively.

For illustrative simplicity, let there be three statements, r_1, r_2, and r_3, that measure three latent traits, θ_1, θ_2, and θ_3, respectively. These three statements produce three pairwise-comparison items: (r_1, r_2), (r_2, r_3), and (r_1, r_3). In the RIM, the log-odds of selecting r_1 over r_2 from the pair (r_1, r_2), selecting r_2 over r_3 from the pair (r_2, r_3), and selecting r_1 over r_3 from the pair (r_1, r_3) are defined, respectively, as:

$$\begin{aligned}
\log\left(\frac{P_{r_1}\left|\left(r_1,r_2\right),n\right.}{P_{r_2}\left|\left(r_1,r_2\right),n\right.}\right) &= \left(\theta_{1n}+u_{r_1}\right)-\left(\theta_{2n}+u_{r_2}\right), \\
\log\left(\frac{P_{r_2}\left|\left(r_2,r_3\right),n\right.}{P_{r_3}\left|\left(r_2,r_3\right),n\right.}\right) &= \left(\theta_{2n}+u_{r_2}\right)-\left(\theta_{3n}+u_{r_3}\right), \\
\log\left(\frac{P_{r_1}\left|\left(r_1,r_3\right),n\right.}{P_{r_3}\left|\left(r_1,r_3\right),n\right.}\right) &= \left(\theta_{1n}+u_{r_1}\right)-\left(\theta_{3n}+u_{r_3}\right),
\end{aligned} \tag{3}$$

where $\theta_{1n}+\theta_{2n}+\theta_{3n}=0$ for every person, and u_{r_1}, u_{r_2}, and u_{r_3} are the utilities for the three statements, respectively.

As the θ variables are centered at zero for every person, they reflect *psychological differentiation* among the latent traits within a person (Witkin, Goodenough, & Oltman 1979). For example, a person with $\theta_1=\theta_2=\theta_3=0$ has no differentiation among the three latent traits; whereas a person with $\theta_1=2$, $\theta_2=1$, and $\theta_3=-3$ has a large differentiation, and among the three latent traits, this person shows a very high level of latent trait θ_1 but a very low level of latent trait θ_3. Psychological differentiation is very important in development (Witkin et al. 1979). For example, differentiation in the Holland occupational themes (Realistic, Investigative, Artistic, Social,, Enterprising, and Conventional) (Holland 1973) is a crucial construct in personality and counseling psychology and is a powerful predictor of attitudes and behaviors in the career decision-making process (Hirschi 2009). Similarly, emotional differentiation (Barrett 2004) plays an adaptive role in daily life (Barrett, Gross, Christensen, & Benvenuto 2001; Kashdan, Barrett, & McKnight 2015), and a low level of differentiation is often associated with psychological disorders (Trull, Lan, Koval, & Ebner-Priemer 2015).

The RIM has several important features (Chen & Wang 2013a). First, the item and person parameters can be separated (i.e., specific objectivity). Second, each statement has a single utility value, no matter with which the statement the RIM is paired. Third, the RIM includes the random utility model (e.g., Luce 1959) as a special case in which the same latent trait is assessed across all statements.

2 Item Response Models for Multidimensional Ranking Items

The RIM is limited to pairwise-comparison items but can be extended to ranking items, to be shown below. To do so, two approaches were adopted in this study, namely, the exploded logit IRT and the generalized logit IRT.

2.1 The Exploded Logit IRT

Let there be K statements $r_1, r_2, \ldots, r_K$ in a ranking item, with utilities of $u_{r_1}, u_{r_2}, \ldots$, and u_{r_K}, respectively. For notational simplicity, we do not include subscripts for the items. Let $r_{(l)}$ be the statement that is given rank l ($l = 1, \ldots, K$), where $l = 1$ represents the first-rank order and $l = K$ represents the Kth-rank order, and let $\mathbf{R} \equiv \{r_{(1)}, r_{(2)}, \ldots, r_{(K)}\}$ be the ranking pattern. For example, three activities are to be ranked: $r_1 =$ attending parties, $r_2 =$ visiting museums, $r_3 =$ solving geometric proofs. If these three activities are ranked as 2, 3, 1, respectively, then the ranking pattern is {solving geometric proofs, attending parties, visiting museums}. To form the exploded logit model, the ranking process is "conceptually" exploded (decomposed) into a series of choices, and the probability of observing the ranking pattern $\mathbf{R}$ is defined as (Allison & Christakis 1994):

$$P(\mathbf{R}) = \prod_{l=1}^{K-1} \frac{\exp\left(u_{r_{(l)}}\right)}{\sum_{m=l}^{K} \exp\left(u_{r_{(m)}}\right)}. \tag{4}$$

As an example, let the three statements r_1, r_2, and r_3 in a ranking item be ranked as 2, 3, and 1, respectively (i.e., $r_{(1)} = r_3$, $r_{(2)} = r_1$, and $r_{(3)} = r_2$). According to Eq. (4), the probability of this ranking pattern is:

$$P(\mathbf{R}) = P(\{r_3, r_1, r_2\}) = \frac{\exp(u_{r_3})}{\exp(u_{r_3}) + \exp(u_{r_1}) + \exp(u_{r_2})} \times \frac{\exp(u_{r_1})}{\exp(u_{r_1}) + \exp(u_{r_2})}. \tag{5}$$

The first term in the right-hand side of Eq. (5) indicates r_3 is first selected among r_1, r_2, and r_3; the second term indicates r_1 is then selected from the remaining r_1 and r_2 because r_3 is selected in the previous choice. Although the ranking process is conceptually decomposed into a series of choices in the exploded logit model, the model does not require respondents to actually choose statements sequentially (Allison & Christakis 1994).

Equation 4 does not involve person measures because all statements are assumed to measure the same latent trait (i.e., unidimensional ranking items). For multidimensional ranking items, we can add person measures $\theta_{r_{(l)},n}$ to Eq. (4):

$$P(\mathbf{R}_n) = \prod_{l=1}^{K-1} \frac{\exp\left(\theta_{r_{(l)},n} + u_{r_{(l)}}\right)}{\sum_{m=l}^{K} \exp\left(\theta_{r_{(m)},n} + u_{r_{(m)}}\right)}, \tag{6}$$

where $P(\mathbf{R}_n)$ is the probability of observing ranking pattern $\mathbf{R}$ for person n, $\theta_{r_{(l)},n}$ is person n's "relative" level of the latent trait that statement $r_{(l)}$ measures, and the others have been defined. As in the RIM, the θ variables are person-centered (i.e., $\sum \theta = 0$ for every person) and assumed to follow a multivariate normal distribution. Equation 6 is referred to as the exploded logit IRT model for multidimensional ranking items (ELIRT). For pairwise-comparison items, the ELIRT reduces to the RIM. That is, the ELIRT includes the RIM as a special case.

2.2 The Generalized Logit IRT

Let a ranking item consist of K statements, r_1, r_2, ..., r_K, measuring $\theta_{r_1}, \theta_{r_2}, \ldots, \theta_{r_K}$, respectively. There are a total of $K!$ possible ranking patterns. Let $\mathbf{\Omega} \equiv \{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_t, \ldots, \mathbf{R}_{K!}\}$ be the collection of all ranking patterns. In the generalized logit IRT model for multidimensional ranking items (GLIRT), the probabilities of these $K!$ ranking patterns $\mathbf{P}_{K!\times 1}(\mathbf{\Omega})$ can be expressed as:

$$\mathbf{P}_{K!\times 1}(\mathbf{\Omega}) = \exp\left(\mathbf{A}_{K!\times K}\mathbf{w}_{K\times 1}\right) / \left(\mathbf{1}_{1\times K!} \exp\left(\mathbf{A}_{K!\times K}\mathbf{w}_{K\times 1}\right)\right), \tag{7}$$

where $\mathbf{A}$ is the design matrix, and $\mathbf{w} = [\theta_{r_1 n} + u_{r_1}, \theta_{r_2 n} + u_{r_2}, \ldots, \theta_{r_K n} + u_{r_K}]^{\mathrm{T}}$. The sequence of $\mathbf{R}_t$ in $\mathbf{\Omega}$ can be arbitrary, but the simplest way is to arrange the set of all permutations by a sequence of transpositions of r_k with its left neighbor. Each row of $\mathbf{A}$ represents a ranking pattern $\mathbf{R}_t$, and the columns within each row provide a set of coefficients for the K statements. Therefore, the design matrix $\mathbf{A}$ can be constructed as follows. Within each row (ranking pattern), the coefficient in a column is equal to $K - l$, where l is the rank order assigned to the corresponding statement.

For illustrative simplicity, let a ranking item consist of three statements r_1, r_2, and r_3 that measure latent traits θ_{r_1}, θ_{r_2}, and θ_{r_3}, respectively. Then, according to the GLIRT, we can have:

$$\begin{aligned}
\mathbf{P}_{6\times1}(\mathbf{\Omega}) &= [P(\{r_1, r_2, r_3\}), P(\{r_1, r_3, r_2\}), P(\{r_3, r_1, r_2\}), P(\{r_3, r_2, r_1\}), \\
&\quad P(\{r_2, r_3, r_1\}), P(\{r_2, r_1, r_3\})]^{\mathrm{T}}. \\
\mathbf{A}_{6\times3} &= \begin{bmatrix} 2\,1\,0 \\ 2\,0\,1 \\ 1\,0\,2 \\ 0\,1\,2 \\ 0\,2\,1 \\ 1\,2\,0 \end{bmatrix} \\
\mathbf{w}_{3\times1} &= [\theta_{r_1n} + u_{r_1}, \theta_{r_2n} + u_{r_2}, \theta_{r_3n} + u_{r_3}]^{\mathrm{T}}.
\end{aligned} \tag{8}$$

Specifically, the probabilities of the six ranking patterns for person n can be expressed as:

$$\begin{aligned}
P(\{r_1, r_2, r_3\}) &= \exp[2(\theta_{r_1n} + u_{r_1}) + (\theta_{r_2n} + u_{r_2})]/\Psi \\
P(\{r_1, r_3, r_2\}) &= \exp[2(\theta_{r_1n} + u_{r_1}) + (\theta_{r_3n} + u_{r_3})]/\Psi \\
P(\{r_3, r_1, r_2\}) &= \exp[2(\theta_{r_3n} + u_{r_3}) + (\theta_{r_1n} + u_{r_1})]/\Psi \\
P(\{r_3, r_2, r_1\}) &= \exp[2(\theta_{r_3n} + u_{r_3}) + (\theta_{r_2n} + u_{r_2})]/\Psi \\
P(\{r_2, r_3, r_1\}) &= \exp[2(\theta_{r_2n} + u_{r_2}) + (\theta_{r_3n} + u_{r_3})]/\Psi \\
P(\{r_2, r_1, r_3\}) &= \exp[2(\theta_{r_2n} + u_{r_2}) + (\theta_{r_1n} + u_{r_1})]/\Psi
\end{aligned} \tag{9}$$

where Ψ is the sum of all numerators in Eq. (9) to make the six probabilities sum to 1, and the others have been defined.

If there are four statements ($K = 4$) in a ranking item, r_1, r_2, r_3, and r_4, that measure latent traits θ_{r_1}, θ_{r_2}, θ_{r_3}, and θ_{r_4}, respectively, then we can have:

$$\begin{aligned}
&\mathbf{P}_{24\times1}(\mathbf{\Omega}) = [P(\{r_1, r_2, r_3, r_4\}), P(\{r_1, r_2, r_4, r_3\}), \ldots, P(\{r_4, r_3, r_1, r_2\}), P(\{r_4, r_3, r_2, r_1\})]^{\mathrm{T}}. \\
&\mathbf{A}_{24\times4} = \begin{bmatrix} 3 & 2 & 1 & 0 \\ 3 & 2 & 0 & 1 \\ \vdots & \ddots & \ddots & \vdots \\ 1 & 0 & 2 & 3 \\ 0 & 1 & 2 & 3 \end{bmatrix} \\
&\mathbf{w}_{4\times1} = [\theta_{r_1n} + u_{r_1}, \theta_{r_2n} + u_{r_2}, \theta_{r_3n} + u_{r_3}, \theta_{r_4n} + u_{r_4}]^{\mathrm{T}}.
\end{aligned} \tag{10}$$

The GLIRT can be generalized to more than four statements and four latent traits. For pairwise-comparison items, the GLIRT reduces to the RIM.

Although the ELIRT and the GLIRT are developed based on different approaches, both models are generalized linear models and have the same number of parameters. Thus, the ELIRT and the GLIRT are expected to fit the ranking data about equally well, although the models' parameters cannot be directly compared because the models are on different scales.

3 Simulations

We conducted simulations to evaluate the parameter recovery of the ELIRT and the GLIRT. There were three latent traits, and each latent trait had seven statements, resulting in a total of 21 statements. The utilities of the 21 statements were generated from $N(0, 1)$, but they summed to zero within each latent trait (for model identification). As shown in Fig. 1, a total of 14 triad items were generated with a linking design to place all statement utilities on a common scale. For example, statements 1, 8, and 15 formed a triad items; statements 1, 9, and 17 formed another triad item. A total of 1000 persons were generated from a multivariate normal distribution with a mean vector of $\mathbf{0}$ and variances of 1, and covariance of -0.5 for θ_1 and θ_2. $\theta_{3n} = -(\theta_{1n} + \theta_{2n})$ because of the ipsative nature. The data were generated from the ELIRT or the GLIRT, and analyzed with the respective data-generating model to evaluate how the parameters could be recovered using Just Another Gibbs Sampler (JAGS; Plummer 2003). Thirty replications were conducted. The priors for the covariance matrix were set as a three-dimensional inverse Wishart distribution $W[\mathbf{I}, k]$ with hyperparameters $k = 2$ and an identity matrix $\mathbf{I}$, the priors for the mean vector of latent traits as $\boldsymbol{\mu} \sim MVN(\mathbf{0}, \mathbf{I})$, and the priors for the statement utilities as $u \sim N(0, 1)$. After 10,000 burn-in iterations, the parameters were estimated based on 1000 draws of iterations from the joint posterior distribution.

Table 1 lists the true, bias, and root mean square error (RMSE) values for the parameters based on the two models. When the ELIRT was fit to the ELIRT data, the RMSE values were between 0.040 and 0.062 for the statement utilities, between 0.044 and 0.064 for the variance–covariance of the latent traits, and between 0.039 and 0.041 for the mean of the latent traits. Similarly, when the GLIRT was fit to the GLIRT data, the parameter recovery was as satisfactory as that in the ELIRT. Specifically, the RMSE values were between 0.034 and 0.067 for the statement utilities, between 0.030 and 0.042 for the variance–covariance of the latent traits, and between 0.046 and 0.054 for the mean of the latent traits. In sum, the parameters in both models were recovered well with JAGS.

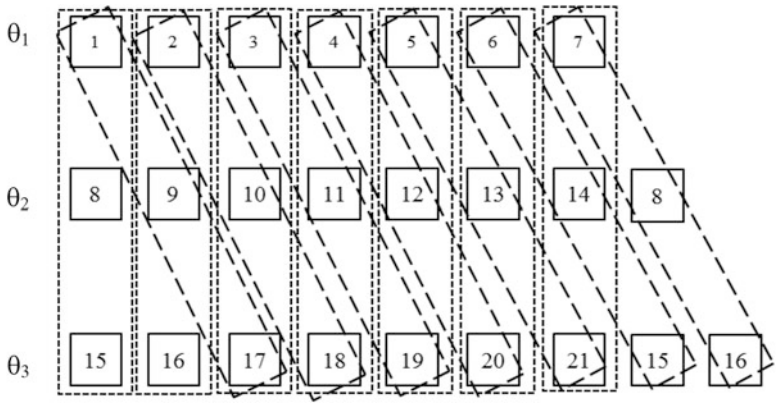


Fig. 1 Linking design of the 21 statements in 14 triad items in the simulation study

Table 1 True, Bias and RMSE values for the parameter estimates in the simulation study under the ELIRT and the GLIRT

	Dim. 1					Dim. 2					Dim.3				
		ELIRT		GLIRT			ELIRT		GLIRT			ELIRT		GLIRT	
Par.	Gen.	Bias	RMSE	Bias	RMSE	Gen.	Bias	RMSE	Gen.	Bias	RMSE	Bias	RMSE		
u_{r_1}	−1.548	0.005	0.054	0.020	0.053	−0.433	0.027	0.051	0.001	0.052	−0.399	0.008	0.045	0.028	0.054
u_{r_2}	−1.230	−0.002	0.056	−0.026	0.067	1.834	−0.004	0.062	−0.028	0.063	−1.376	0.001	0.060	−0.015	0.060
u_{r_3}	−0.001	0.009	0.040	−0.004	0.034	0.691	−0.015	0.044	−0.004	0.046	0.146	−0.010	0.051	−0.001	0.034
u_{r_4}	1.197	−0.013	0.050	−0.019	0.047	0.687	−0.013	0.050	−0.024	0.051	1.178	−0.010	0.042	−0.016	0.055
u_{r_5}	0.298	−0.001	0.051	0.015	0.055	−0.335	−0.015	0.049	0.005	0.046	1.100	0.003	0.046	0.006	0.050
u_{r_6}	0.930	−0.006	0.052	0.006	0.050	−0.981	0.011	0.051	0.028	0.056	1.634	−0.007	0.062	−0.012	0.046
u_{r_7}	0.355	0.006	0.058	0.007	0.046	−1.463	0.009	0.051	0.021	0.057	−2.282	0.013	0.049	0.009	0.062
μ	0.000	0.006	0.039	−0.007	0.054	0.000	−0.005	0.041	−0.011	0.046	0.000	−0.001	0.039	0.014	0.053
σ^2	1.000	0.007	0.064	−0.005	0.030	1.000	0.022	0.065	−0.009	0.032	1.000	0.037	0.063	0.014	0.031
$\sigma_1\sigma_2$	−0.500	0.004	0.044	0.016	0.039										
$\sigma_1\sigma_3$	−0.500	−0.011	0.046	−0.009	0.042										
$\sigma_2\sigma_3$	−0.500	−0.026	0.048	−0.005	0.039										

Note: u_{r_1} to u_{r_7} are statement utilities
μ is the latent trait mean
σ^2 is the latent trait variance
$\sigma_1\sigma_2$ to $\sigma_2\sigma_3$ describe the covariance between latent traits
Par. parameter, *Gen.* generating values, *ELIRT* exploded logit IRT model for multidimensional ranking items, *GLIRT* generalized logit IRT model for multidimensional ranking items, *RMSE* root mean square error

4 An Empirical Example

The data included 66 triads with 132 statements, measuring 12 latent traits: Energy, Assertiveness, Sociability, Concern for Others, Dependability, Organized, Achievement Orientation, Initiative, Multitasking, Innovative, Self-Confidence, and Self-Control. Each latent trait was measured with 11 statements. Among the 66 triads, 22 had statements that were also shown in other triads to place all 132 statements on the same scale for comparison. A total of 1675 respondents completed the questionnaire. Of the respondents, approximately 845 (50.4 %) were White, 296 (17.7 %) were African American, 355 (21.2 %) were Asian, 94 (5.6 %) were Hispanic, and 85 (5.1 %) were another ethnicity; about half of the respondents were women. More than 95 % of the respondents completed the test without missing data. The respondents were informed that the testing was for scale development to understand how people behave at workplaces, and they were instructed to rank the statements based on their preferences. A sample item consisted of three statements: (a) I stand up for my rights, (b) When solving a problem, I strive to discover the very best solution, and (c) If something needs to be done, I get it done without having to be told. These three statements measured Assertiveness, Dependability, and Initiative, respectively. The first-, second-, and third-rank orders were scored as 2, 1, and 0, respectively.

Columns 2 and 3 of Table 2 show the descriptive statistics of the raw scores for each dimension of the test. For each respondent, the raw score of a latent trait was calculated by averaging his or her responses to all the statements measuring that latent trait. The higher the raw score, the higher the ranking of the latent trait. Across all respondents, Initiative was ranked the highest ($M = 1.09$, $SD = 0.23$), followed by Self-Confidence ($M = 1.06$, $SD = 0.29$), whereas Self-Control was the lowest ($M = 0.82$, $SD = 0.27$).

We then fit the ELIRT and the GLIRT to the data using JAGS. The specifications of the priors were similar to those of the simulation studies. After 15,000 burn-in iterations, convergence was achieved, and the parameters were estimated based on 1000 draws from the joint posterior distribution. It took approximately 40 h to converge for both models. These two models were compared according to the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linden 2002). The model-data fit was checked by using the posterior predictive model checking (PPMC) method (Gelman, Meng & Stern 1996) with the frequency of each ranking pattern across items as the discrepancy measure.

The results showed that the GLIRT had a smaller DIC (359,493) than the ELIRT (360,024), suggesting that the GLIRT was preferred. The posterior predictive p values were between .28 and .68 ($M = .47$, $SD = 0.16$) for the GLIRT and between .029 and .975 ($M = .53$, $SD = 0.38$) for the ELIRT, suggesting both models had a good fit. The two models yielded very similar results. For example, the estimates for the statement utilities ($r = .93$), the means of the latent traits ($r = .98$), and the variance–covariance matrix ($r = .99$) were highly correlated between the models. According to the GLIRT, the statement utilities measuring the 12 latent traits (mean

Table 2 Means and standard deviations of raw scores, and estimates and standard errors for the mean levels of the 12 latent traits obtained from the GLIRT in the empirical example

Latent trait	Mean	SD	Est.	SE
1. Energy	0.97	0.30	0.04	0.014
2. Assertiveness	0.97	0.31	−0.04	0.019
3. Sociability	0.99	0.32	0.00	0.017
4. Concern for others	0.99	0.26	−0.01	0.015
5. Dependability	1.02	0.24	0.02	0.011
6. Organized	1.03	0.28	−0.05	0.017
7. Achievement orientation	1.03	0.23	0.03	0.014
8. Initiative	1.09	0.23	0.08	0.022
9. Multitasking	1.02	0.26	0.03	0.013
10. Innovative	1.01	0.26	0.06	0.013
11. Self-confidence	1.06	0.29	0.07	0.014
12. Self-control	0.82	0.27	−0.22	NA

Note: *Est.* estimates, *NA* denotes the standard error was not applicable because the 12th latent trait was constrained, *GLIRT* generalized logit IRT model for multidimensional ranking data

zero for every latent trait) ranged from −0.232 to 0.361, −0.342 to 0.562, −0.487 to 0.760, −0.757 to 0.625, −0.946 to 0.803, −0.618 to 0.416, −0.743 to 0.581, −0.986 to 0.369, −0.663 to 0.455, −0.531 to 0.732, −0.612 to 0.394, and −0.658 to 0.593, respectively. Among the 132 statements, statement 138 measuring Initiative had the lowest utility (–0.986) while statement 257 measuring Dependability had the highest utility (0.803). Columns 4 and 5 of Table 2 show the means for the 12 latent traits across respondents and their standard errors, respectively. Initiative was the highest ($M = 0.08$, SE = 0.022), followed by Self-Confidence ($M = 0.07$, $SE = 0.014$), whereas Self-Control was the lowest ($M = -0.22$, SE is not available because the estimate was constrained as $\theta_{12} = -\sum_{d=1}^{11} \theta_d$ for every person). These results were consistent with the mean raw scores shown in column 2 of Table 2, with a Pearson correlation between the raw score means and the latent trait means of .88. The correlations among the latent traits, shown in Table 3, were mainly negative and moderate. Innovative and Self-Confidence had the highest positive correlation of .617; Assertiveness and Self-Confidence had the highest negative correlation of −.546. Although the θ variables were negatively correlated, it did not necessarily mean that the η variables were also negatively correlated. When the η variables are uncorrelated, the expected correlations between ipsative measures equals $-1/(D-1)$, where D is the number of latent traits (Aitchison 1986, pp. 52–62; Dunlap & Cornwell 1994).

In terms of the individual difference in the latent traits, the person measures of the latent traits can be estimated from the means of the posterior distributions. Table 4 and Fig. 2 present the latent trait measures of two example persons. For Person 1, among the 12 latent traits, Sociability was the highest (0.83) and Self-Control was the lowest (−0.63), suggesting this person valued Sociability more than Self-Control. For Person 2, Energy was the highest (0.65) and Self-Control was the lowest (−0.27), suggesting this person valued Energy more than Self-Control.

Table 3 Variance-covariance and correlation matrices among the 12 latent traits under the GLIRT in the empirical example

Latent trait	1	2	3	4	5	6	7	8	9	10	11	12
1	0.113	0.034	0.047	−0.017	−0.033	−0.029	−0.007	−0.009	−0.026	−0.030	−0.011	−0.031
2	0.255	0.156	0.059	0.004	−0.032	−0.010	−0.025	−0.026	−0.043	−0.053	−0.058	−0.006
3	0.359	0.385	0.149	0.020	−0.031	−0.060	−0.032	−0.016	−0.007	−0.052	−0.048	−0.027
4	−0.220	0.045	0.226	0.054	0.002	0.002	−0.013	−0.010	0.007	−0.017	−0.026	−0.005
5	−0.531	−0.440	−0.448	0.040	0.033	0.008	0.006	0.012	0.015	0.015	0.007	−0.002
6	−0.285	−0.087	−0.515	0.032	0.137	0.091	0.007	−0.003	−0.005	−0.001	−0.002	0.002
7	−0.124	−0.378	−0.493	−0.329	0.209	0.136	0.029	0.008	0.005	0.012	0.019	−0.008
8	−0.135	−0.333	−0.211	−0.215	0.333	−0.045	0.225	0.039	0.011	0.010	0.007	−0.022
9	−0.376	−0.531	−0.094	0.140	0.409	−0.077	0.139	0.265	0.041	0.006	0.006	−0.010
10	−0.348	−0.520	−0.517	−0.289	0.319	−0.013	0.276	0.186	0.115	0.067	0.043	0.001
11	−0.128	−0.546	−0.465	−0.427	0.139	−0.029	0.414	0.134	0.113	0.617	0.071	−0.007
12	−0.275	−0.043	−0.210	−0.062	−0.034	0.023	−0.138	−0.333	−0.147	0.014	−0.076	0.115

Note: 1 Energy, 2 Assertiveness, 3 Sociability, 4 Concern for others, 5 Dependability, 6 Organized, 7 Achievement orientation, 8 Initiative, 9 Multitasking, 10 Innovative, 11 Self-confidence, 12 Self-control, GLIRT generalized logit IRT model for multidimensional ranking items
The variance-covariance are in the *upper triangle* and the correlations are in the *lower triangle*

Table 4 Measures of the 12 latent traits for two sample persons under the GLIRT in the empirical example

	Latent trait												
Person	1	2	3	4	5	6	7	8	9	10	11	12	*SD*
#1	0.49	0.38	0.83	0.09	−0.25	−0.3	−0.08	−0.07	−0.02	−0.21	−0.23	−0.63	0.39
#2	0.65	0.12	0.48	−0.25	−0.22	−0.22	−0.08	−0.01	−0.05	−0.09	−0.06	−0.27	0.29

Note: 1 Energy, 2 Assertiveness, 3 Sociability, 4 Concern for others, 5 Dependability, 6 Organized, 7 Achievement orientation, 8 Initiative, 9 Multitasking, 10 Innovative, 11 Self-confidence, 12 Self-control, GLIRT Generalized logit IRT model for multidimensional ranking items

The person measures of Energy were 0.49 and 0.65 for Persons 1 and 2, respectively, indicating Person 2 had a higher differentiation on Energy among the latent traits than Person 1. In contrast, the person measures of Sociability were 0.83 and 0.48 for Persons 1 and 2, respectively, suggesting Person 1 had a higher differentiation on Sociability among the latent traits than Person 2. Similar comparisons can be conducted on the other latent traits. Moreover, the variances of the person measures were 0.156 and 0.084 for Persons 1 and 2, respectively, indicating that Person 1 had a higher differentiation on the 12 latent traits than Person 2.

5 Summary and Discussion

We have successfully developed two new IRT models for multidimensional ranking items. Statement utilities and person estimates are available in the proposed models. Because of the indeterminacy of the scale origin in ipsative data, the absolute levels of latent traits (the η variables) are not identifiable, but the relative levels of latent traits (the θ variables) are identifiable with proper constraints. When focusing on the θ variables that reflect differentiation among latent traits, the variables can be used for interindividual and intraindividual comparison. Simulation studies demonstrated that the parameters of both models could be well recovered using the freeware JAGS, and the empirical example illustrates their applications. Both models appear to work properly, and users are free to apply either one.

We limited the number of conditions in the simulation studies mainly because each replication can take many hours for parameter convergence. Future simulation studies can be conducted under more comprehensive conditions, such as different numbers of dimensions, different numbers of statements per item or per dimension, and different patterns of linking design. Applications of the new models to other empirical datasets are also desirable. In the ELIRT and the GLIRT, a statement has a single utility across all respondents. In practice, a statement may exhibit different degrees of utility for different groups of respondents, which is referred to as differential statement functioning (Chen & Wang 2014). A routine check on differential statement functioning should be part of the development process of ranking tests to ensure test fairness and validity. Computerized adaptive testing is another challenging topic. How to adapt current computerized adaptive testing algorithms to multidimensional ranking items requires further investigation. Although the authors have conducted pioneering work on differential statement functioning and computerized adaptive testing for multidimensional pairwise-comparison and ranking items (Chen & Wang 2013b 2014; Chen, Wang & Ro 2015a; 2015b; Qiu & Wang 2014; Qiu, Wang & Ro 2015), more research is needed. In practice, partial rankings and ties are often found in ranking items (Skrondal & Rabe-Hesketh 2003). Recently, it has become popular to add covariates to IRT models to explain variations in person measures or item parameters, which calls for explanatory IRT (De Boeck & Wilson 2004). Future studies could be conductedto investigate these issues.

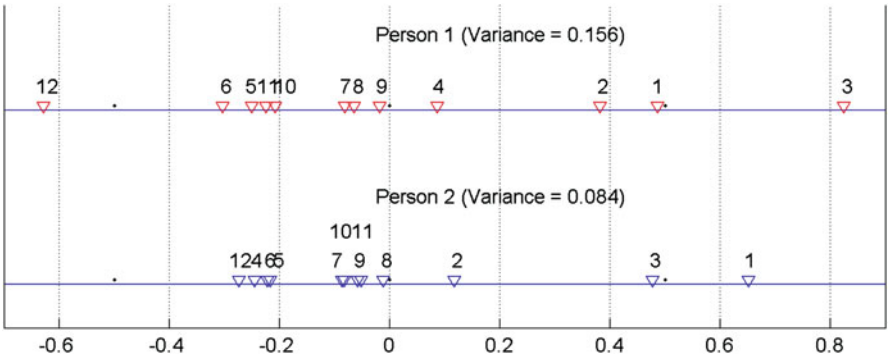


Fig. 2 Measures of the 12 latent traits for two sample persons under the GLIRT in the empirical example. *Note*: (1) Energy; (2) Assertiveness; (3) Sociability; (4) Concern for others; (5) Dependability; (6) Organized; (7) Achievement Orientation; (8) Initiative; (9) Multitasking; (10) Innovative; (11) Self-Confidence; (12) Self-Control. GLIRT generalized logit IRT model for multidimensional ranking items

Acknowledgement The study was supported by the General Research Fund, Hong Kong Research Grants Council (No. 845013).

References

Aitchison, J. A. (1986). *The statistical analysis of compositional data*. London, UK: The Blackburn Press.

Allison, P. D., & Christakis, N. A. (1994). Logit models for sets of ranked items. *Sociological Methodology, 24*, 199–228.

Barrett, L. F. (2004). Feelings or words? Understanding the content in self-report ratings of experienced emotion. *Journal of Personality and Social Psychology, 87*, 266–281.

Barrett, L. F., Gross, J., Christensen, T. C., & Benvenuto, M. (2001). Knowing what you are feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation. *Cognition and Emotion, 15*, 713–724.

Beggs, S., Cardell, S., & Hausman, J. (1981). Assessing the potential demand for electric cars. *Journal of Econometrics, 17*, 1–19.

Böckenholt, U. (2004). Comparative judgments as an alternative to ratings: Identifying the scale origin. *Psychological Methods, 9*, 453–465.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs I: The method of paired comparisons. *Biometrika, 39*, 324–345.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*, 460–502.

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using *Mplus. Behavioral Research Methods, 44*, 1135–1147.

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18*, 36–52.

Cattell, R. B. (1944). Psychological measurement: Normative, ipsative, interactive. *Psychological Review, 51*, 292–303.

Chapman, R. G., & Staelin, R. (1982). Exploiting rank ordered choice set data within the stochastic utility model. *Journal of Marketing Research, 14*, 288–301.

Chen, C.-W., & Wang, W.-C. (2013a, April). *Item response theory models for ipsative tests*. Paper presented at the annual meeting of National Council on Measurement in Education, San Francisco, CA.

Chen, C.-W., & Wang, W.-C. (2013b, July). *Computerized adaptive testing under the Rasch model for ipsative forced-choice items*. Paper presented at the 78th annual meeting of the Psychometric Society, Arnhem, Netherlands.

Chen, C.-W., & Wang, W.-C. (2014, April). *Detecting differential statement functioning in ipsative tests using the logistic regression method*. Paper presented at the annual meeting of National Council on Measurement in Education, Philadelphia, PA.

Chen, C.-W., Wang, W.-C., & Ro, S. (2015a, July). *A quick item selection method in computerized adaptive testing for ranking items*. Paper presented at the 80th annual meeting of the Psychometric Society, Beijing, China.

Chen, C.-W., Wang, W.-C., & Ro, S. (2015b, August). *Controlling within-person exposure in computerized adaptive testing for ranking items*. Paper presented at the Pacific Rim Objective Measurement Symposium, Fukuoka, Japan.

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

de Vries, A. L. M., & van der Ark, A. (2008). Scoring methods for ordinal multidimensional forced-choice items. In J. Daunis-i-Estadella, & J. A. Martín-Fernández (Eds.), *Proceedings of the 3rd Compositional Data Analysis Workshop Codawork '08* (pp. 1–18). Girona, Spain: University of Girona.

Dunlap, W. P., & Cornwell, J. M. (1994). Factor analysis of ipsative measures. *Multivariate Behavioral Research, 29*, 115–126.

Gelman, A., Meng, X.-L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica, 6*, 733–807.

Hausman, J. A., & Ruud, P. A. (1987). Specifying and testing econometric models for ranked-ordered data. *Journal of Econometrics, 34*, 83–104.

Hirschi, A. (2009). Development and criterion validity of differentiated and elevated vocational interests in adolescence. *Journal of Career Assessment, 17*, 384–401.

Holland, J. L. (1973). *Making vocational choices: A theory of careers*. Englewood Cliffs, NJ: Prentice-Hall.

Kashdan, T. B., Barrett, L. F., & McKnight, P. E. (2015). Unpacking emotion differentiation: Transforming unpleasant experience by perceiving distinctions in negativity. *Current Direction in Psychological Science, 24*, 10–16.

Kopelman, R. E., Rovenpor, J. L., & Guan, M. (2003). The study of values: Construction of the fourth edition. *Journal of Vocational Behavior, 62*, 203–220.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York, NY: Wiley.

Maydeu-Olivares, A., & Böckenholt, U. (2005). Structural equation modeling of paired-comparison and ranking data. *Psychological Methods, 10*, 285–304.

Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research, 45*, 935–974.

Meade, A. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology, 77*, 531–552.

Mokken, R. J. (1971). *A theory and procedure of scale analysis with applications in political research*. Berlin: Walter de Gruyter, Mouton.

Plummer, M. (2003, March). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. Paper presented at the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria.

Punj, G. N., & Staelin, R. (1978). The choice process for graduate business schools. *Journal of Marketing Research, 15*, 588–598.

Qiu, X.-L., & Wang, W.-C. (2014, April). *Computerized adaptive testing for forced-choice ipsative items*. Paper presented at the annual meeting of American Educational Research Association, Philadelphia, PA.

Qiu, X.-L., Wang, W.-C., & Ro, S. (2015, April). *An IRT model for multidimensional ranking data in ipsative tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Salgado, J. F., & Tauriz, G. (2014). The five-factor model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology, 23*, 3–30.

Saville, P., Sik, G., Nyfield, G., Hackston, J., & MacIver, R. (1996). A demonstration of the validity of the Occupational Personality Questionnaire (OPQ) in the measurement of job competencies across time and in separate organizations. *Applied Psychology, 45*, 243–262.

SHL. (2006). *OPQ32 technical manual*. Thames Ditton: SHL Group. Retrieved from https://www.cebglobal.com/shl/uk/solutions/products/docs/OPQ_Fact_Sheet_CEB%20v1.pdf.

Skrondal, A., & Rabe-Hesketh, S. (2003). Multilevel logistic regression for polytomous data and rankings. *Psychometrika, 68*, 267–287.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linden, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B, 64*, 583–640.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 79*, 281–299.

Thurstone, L. L. (1931). Rank order as a psychological method. *Journal of Experimental Psychology, 14*, 187–201.

Trull, T. T., Lan, S. P., Koval, P., & Ebner-Priemer, U. W. (2015). Affective dynamics in psychopathology. *Emotion Review, 7*, 355–361.

Witkin, H. A., Goodenough, D. R., & Oltman, P. K. (1979). Psychological differentiation: Current status. *Journal of Personality and Social Psychology, 37*, 1127–1145.

Different Growth Measures on Different Vertical Scales

Dongmei Li

Abstract Vertical scales have been used by testing programs for decades to facilitate the tracking of student performance over time. With the recent emphasis on the measuring of student growth for accountability purposes, scores from vertically scaled tests have been used to evaluate school or teacher performance. Because there are different types of growth measures and there are also different ways to construct a vertical scale, it is important to understand the impact of the vertical scales on various growth measures for important educational decisions. Based on some mathematical relationships that have been shown to exist among certain growth measures and with the use of empirical data, this study investigated the impact of different vertical scales on the relationships among simple gain scores, residual gain scores, and three growth measures based on conditional status percentile ranks (CSPR). Results showed that the correlations between simple gain scores and the rest of the growth measures were affected by the extent of scale expansion or scale shrinkage across grades.

Keywords Growth measures • Vertical scales • Gain scores • Student growth percentiles

Vertical scales are constructed by linking scores of tests from different grade levels through various data collection designs and statistical methods to a common scale. It has been well-acknowledged in the literature that different choices in the vertical scaling process may result in different scales (Li, 2015a; Paek, Young & Yi 2008; Patz 2007; Tong & Kolen 2007). With the increasing high-stakes use of growth measures based on vertically scaled scores in state accountability systems, the effect of different potential vertical scales on growth measures has been a topic of continuous investigation. Research has shown that the rank ordering of students or schools may change not only with the change of growth measures (Dunn & Allen 2009; Goldschmidt, Choi & Beaudon 2012; Li & Kolen 2008), but also with the change of the underlying vertical scales (Briggs & Domingue 2013;

D. Li (✉)
ACT, Inc., 500 ACT Drive, Iowa City, IA 52243, USA
e-mail: dongmei.li@act.org

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer
Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_6

Briggs & Weeks 2009; Lei & Zhao 2012; Li & Kolen 2011). Based on some mathematical relationships that have been shown to exist among certain growth measures, this study sought to investigate, in some greater depth, the effect of different vertical scales on the relationships among a few growth measures. This study also investigated whether certain types of vertical scales tended to yield more consistent results across growth measures than other types of vertical scales. Specifically, this study investigated how certain features of the vertical scales might impact the relationships among three types of growth measures: simple gain scores, residual gain scores, and CSPR.

In the sections below, an introduction is first given on the potential differences among vertical scales. Then, a few growth measures and their known relationships are described. After that, theoretical and empirical investigation results on the impact of vertical scales on the relationships among these measures are described, followed by conclusions and discussion.

1 Properties of Vertical Scales

Strictly speaking, there has been no formal classification of vertical scales into different types in the literature. Vertical scales are often described or compared in terms of grade to grade means, grade to grade score variabilities, or the effect sizes of between-grade mean changes.

Among these features, the differences in grade to grade score variability have probably drawn the most attention in the literature (e.g., Burket 1984; Hoover 1984; Phillips & Clarizio 1988). A quantity depicting the change of score variability across grades, that is, the ratio of the standard deviations of scores for the upper and lower grades, has been shown to be predictive of correlations among a few growth measures (Li 2015b; Li & Kolen 2011; Roberts & Burrill 1995). Relevant to the purpose of this paper is the mathematical relationship between simple gain scores and residual gain scores given by Li and Kolen (2011). This relationship is discussed in the next section.

2 A Few Growth Measures and Their Relationships

This study investigated three types of growth measures that are commonly used in the context of measuring student growth between years: simple gain scores, residual gain scores, and CSPR. For the simplicity of discussion, this study focused on two years' scores, referred to as the current year and the the previous year scores.

The simple gain score is the difference between scores earned between two years. The residual gain score is the difference between a student's observed current year score and the student's predicted current year score, usually based

on a linear regression of the current year scores on the previous year scores. CSPR is a term suggested by Castellano and Ho (2013) to describe the use of percentile ranks of students' current status scores conditioning on previous scores as a growth measure. One example of CSPR is the well-known student growth percentiles (SGPs) (Betebenner 2008, 2009) which have been used in many states' accountability systems. SGPs and two other less commonly used CSPR alternatives were considered in this paper: the ordinary least squares percentile ranks of residuals (PRRs) (Castellano & Ho 2013), and the empirical conditional percentile ranks (ECPRs).

Before investigating how certain features of vertical scales may affect the relationships among these growth measures, some of their known relationships from previous research are described below.

2.1 Simple Gain and Residual Gain Scores

Whether simple gain scores or residual gain scores should be used in measuring change has been debated in the literature (e.g. Harris 1963; Maris 1998). Lord's paradox (Lord 1967) described the inconsistency of results when the evaluation was based on simple gain scores or residual gain scores. Li and Kolen (2011) showed that correlations between simple gain scores and residual gain scores are determined by two quantities: the correlation between time 1 and time 2 scores, and the ratio of standard deviations (SDs) between time 2 and time 1 scores. Let $k = \frac{\sigma_{X_2}}{\sigma_{X_1}}$, where X_1 and X_2 are the time 1 and time 2 scores, and let σ_{X_1} and σ_{X_2} represent the SDs of the time 1 and time 2 scores, respectively. Let $\rho_{X_1X_2}$ represent the correlation between time 1 and time 2 scores. The correlation between simple gain and residual gain scores (ρ_{DR}, where R stands for the residual gain score, and D stands for the simple gain score) is determined by the values of k and $\rho_{X_1X_2}$. Specifically,

$$\rho_{DR} = \frac{k\sqrt{1-\rho_{X_1X_2}^2}}{\sqrt{1+k^2-2k\rho_{X_1X_2}}}, \tag{1}$$

where $R = X_2 - \widehat{X}_2$. The predicted time 2 score $\widehat{X}_2$ is based on linear regression of X_2 on X_1. Li and Kolen (2011) pointed out that Eq. (1) could be used to predict the consistency of results when these two measures were used for evaluating school effectiveness, and that it could also be helpful for understanding in what situations Lord's paradox (Lord 1967) tended to be more severe. This paper applied this equation in the context of vertical scaling and showed its merit in predicting the impact of vertical scales on the relationships between simple gain scores and residual gain scores.

2.2 Three CSPR Measures

As mentioned earlier, this study considered three approaches for estimating growth using CSPR. One was the SGP statistic that has been implemented in many state accountability systems. The other two (PRR and ECPR) were alternatives that have been mainly investigated in research studies.

SGPs are the percentile ranks of current scores conditioning on prior scores based on quantile regression. PRRs, percentile ranks of residuals based on ordinary least square regression, are the percentile ranks of residual gain scores. In addition to the regression methods, one major difference between SGPs and PRRs is that the SGP is the percentile rank in the conditional distributions, but the PRR is the non conditional percentile rank of all residual scores. In other words, SGPs compare a student's current score with the current scores of those who had the same previous scores, whereas PRRs compare a student with all other students in terms of the difference between their current scores and their own expected scores, with the expected scores defined by the linear regression of current year scores on previous scores. Despite the above conceptual differences, it is expected that these two approaches would give similar results as long as the bivariate distribution of the two year scores satisfies the assumptions of a linear relationship and homoscedasticity. This expectation was confirmed by Castellano and Ho (2013). They compared SGPs and PRRs in great depth and concluded that these two metrics were very similar in practice. Using simulations, they also found that PRRs out performed SGPs under multivariate normal distributions (MVN), but SGPs outperformed PRRs with greater deviation from MVN and with greater non linear transformations of the score scales.

ECPRs are conceptually very close to SGPs, except that the ECPR estimates are not based on conditional percentile ranks from quantile regression, but based on the empirical cumulative frequency distributions of current status scores conditioning on prior scores. Some research used these empirical percentile ranks from large data sets as the criteria for evaluating SGPs (e.g., Grady, Lewis & Gao 2010), but the author of this paper considered ECPRs as an alternative approach to SGPs when sufficiently large data are available. Due to the conceptual closeness between ECPRs and SGPs, it was expected that these two approaches would give similar estimates, especially when sample sizes are large.

2.3 Scale Dependency of the Growth Measures

Note that among these growth measures, only simple gain scores directly measure the magnitude of growth between grades and thus require the availability of a vertical scale for their proper use. The other two are based on the conditional distributions and can be used regardless of whether the scores are vertically scaled or not. As confirmed by some studies (Briggs & Domingue 2013; Li 2015b), growth

measures that do not require a vertical scale should be unaffected or at least affected less by changes in vertical scales than growth measures that do require a vertical scale. However, little can be found in the literature regarding how certain features of the vertical scale may affect the relationships among different growth measures. The next section shows how Eq. (1) can be used for this purpose.

3 Vertical Scales and the Relationships Among the Growth Measures

Equation (1) shows that the correlation between simple gain scores and residual gain scores is related to two quantities, the ratio of standard deviations between the upper grade and lower grade (k) and the correlations of scores between the two grades ($\rho_{X_1X_2}$). The quantity k depicts the change of score variability across grades—one important feature of vertical scales. With $k > 1$, the variability of scores increases from the lower grade to the higher grade; with $k = 1$, the variability stays constant; with $k < 1$, the variability decreases, which is often referred to as scale shrinkage. Though the values of the correlations may not be known before longitudinal data are gathered, a few likely values can be used for calculation.

Kolen and Brennan (2014) described several statistical methods that could be used to construct vertical scales, among which the IRT method and the Thurstone (1938) method are the most commonly used. The Thurstone method establishes linkages across grades through linear transformations of z-scores, and the IRT method establishes linkages across grades through linear transformation of the theta scores. Unless there is some other transformation, all variations of scales using the same method (Thurstone or IRT) are linearly related. For those vertical scales that are linearly related, the correlation of scores between grades remains the same no matter which vertical scale is used. In this case, the change of correlations between simple gain and residual gain scores from one vertical scale to another only depends on k.

Figure 1 shows the calculated correlations between simple gain and residual gain scores for a selected range of k and a few between-grade correlation values that are realistic in educational tests based on a review of the literature. This plot shows that the correlations between simple gain and residual gain scores may vary from below .5 (when the scale has a severe shrinkage from lower to upper grade) to well above .9 (when the scale expands in variability to a certain extent). However, the change of correlations between simple gain and residual gain scores also depends on the correlations between scores from the two grades—a greater extent of scale expansion may not necessarily result in a higher correlation between the two growth measures. When score variability stays constant across grades ($k = 1$), the correlations between simple gain and residual gain scores are around .9 or higher for the selected between-grade correlations (.6, .7, .8, and .9).

Equation (1) can be used to predict the correlations of simple gain scores and residual gain scores based on two year data, but what about their relationships with

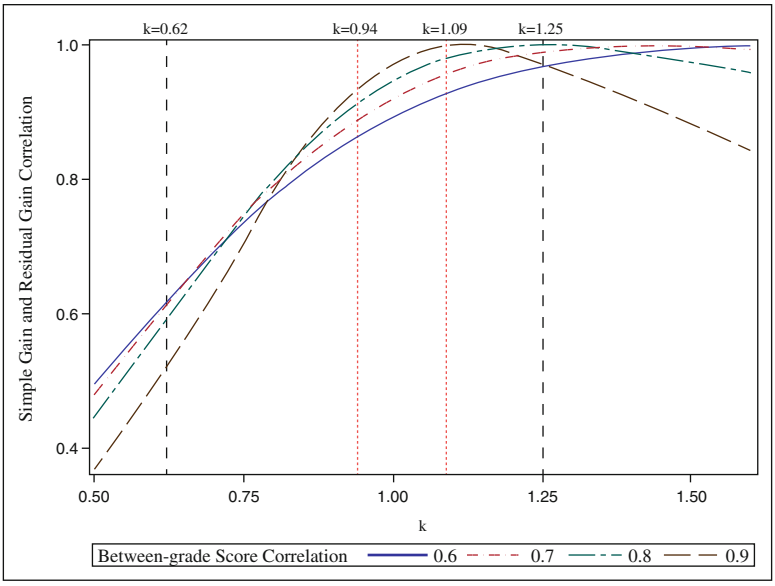


Fig. 1 Correlation of simple gain scores and residual gain scores for different between-grade score correlations and different values of *k*

the CSPR measures? Based on the known relationships between these measures, the following hypotheses are made.

1. Residual gain scores and PRRs are non linearly perfectly related because PRRs are just the percentile ranks of residual gain scores. The observed correlations between them could be less than perfect because of non linearity of their relationship but may still be very high.
2. The high correlations among the three CSPR measures and the perfect relationship between residual gain scores and PRRs may indicate that residual gain scores and other CSPR measures would also be highly correlated. Since none of these four measures require a vertical scale, the high correlations among them are not likely to be affected by the change of the vertical scale.
3. The high correlations among the three CSPR measures and the expected high correlations between residual gain scores and CSPR measures suggest that the correlations between simple gain scores and CSPRs may be similar to the correlations between simple gain scores and residual gain scores. For example, if the correlations between simple gain scores and residual gain scores are low, then it is expected that the correlations between simple gain scores and the CSPR measures would also be low.

Therefore, though residual gain scores and CSPRs do not require vertical scores, when these measures are based on scores from vertical scales, the extent of scale expansion or scale shrinkage of the vertical scale is expected to affect the corre-

lations between these growth measures and simple gain scores. The relationship revealed in Eq. (1) could be used not only to predict the correlations between simple gain and residual gain scores, but also to roughly predict the correlations between simple gain scores and the CSPR measures.

Below are two examples from the literature showing the variability of vertical scales that could result from different choices in the scaling process. These examples are used to demonstrate how Eq. (1) could be used to predict the relationships between simple gain scores and residual gain scores as well as their relationships with the CSPR measures.

3.1 Two Vertical Scaling Examples

Li (2015a) described the results of various scaling options using the Thurstone (1938) vertical scaling methodology, including different linking approaches, different data collection designs, and different variations within the Thurstone method. The study was based on simulated data mimicking the real vertical scaling data collection designs of a large scale testing program for grades 3 through 10. The k values of between-grade SD ratio calculated from the SDs provided in that study ranged from 0.94 to 1.09 across grade pairs and across different scaling options. This range of k values is represented by the two dotted vertical lines in Fig. 1. Correlations between simple gain scores and residual gain scores are expected to be high (above .8) for this set of vertical scales.

Another example came from Briggs and Domingue (2013). They discussed a few vertical scales constructed from empirical data using the IRT methods. Calculated from the SDs provided in their paper, the k values between adjacent grades across the different vertical scales ranged from 0.62 to 1.25, as indicated by the two dashed vertical lines in Fig. 1. Unlike the set of vertical scales from Example 1, this set of vertical scales were more variable and the consistency of rank orders based on simple gain and residual gain scores could vary a lot among these scales.

These two examples show how Eq. (1) can be used for predicting the consistency of growth measures based on simple gain and residual gain scores by the features of the vertical scales even at the stage of scale construction before longitudinal data are available. Hypotheses (1) through (3) stated previously can be applied to predict the relationships among other growth measures, but these hypotheses cannot be tested with the statistics reported in the two example research studies. The next section describes the use of empirical data in examining the relationships among all the growth measures: simple gain scores, residual gain scores, and the three CSPR measures.

3.2 Empirical Data Comparison

A longitudinal data set containing English test scores of about 240,000 students from two different grades was used for the analysis. The operational scale scores (Scale 1) of the two grades had similar standard deviations ($k = 1.024$) and the correlation of scores between grades was around .73. To demonstrate the impact of vertical scales, the operational scale scores were manipulated through linearly transforming the upper grade scores into scores on two fake scales. One manipulation resulted in a vertical scale (Scale 2) with scale expansion ($k = 1.638$) and the other resulted in a vertical scale (Scale 3) with scale shrinkage ($k = 0.617$). Simple gain scores, residual gain scores, SGPs, PRRs, and ECPRs were calculated for each student. Then the correlations among these growth measures were calculated and their scatter plots were examined.

Table 1 shows the correlations of the selected growth measures on the operational scale and on the two manipulated scales. The correlations between simple gain scores and residual gain scores varied depending on the score variability across grades. The correlation between these two measures was high (.94) on the operational scale (Scale 1) on which score variability were similar across grades. On the scale with the larger scale expansion (Scale 2), the correlation was even higher (.99), but on the scale with scale shrinkage (Scale 3), the correlation decreased to .61. These correlations indicated that the rank ordering of students or schools based on simple gain scores or residual gain scores were expected to be mostly consistent on the first two scales, but expected to vary considerably on Scale 3. All these observed correlations were consistent with what would have been expected by applying Eq.

Table 1 Observed correlations among five growth measures on three alternative scales

	Gain	Residual	ECPR	PRR	SGP
Scale 1 (Operational, $k = 1.024$)					
Gain	1.00	0.94	0.90	0.90	0.90
Residual	0.94	1.00	0.95	0.96	0.96
ECPR	0.90	0.95	1.00	0.99	0.99
PRR	0.90	0.96	0.99	1.00	0.99
SGP	0.90	0.96	0.99	0.99	1.00
Scale 2 ($k = 1.638$)					
Gain	1.00	0.99	0.94	0.95	0.94
Residual	0.99	1.00	0.95	0.96	0.95
ECPR	0.94	0.95	1.00	0.99	0.99
PRR	0.95	0.96	0.99	1.00	0.99
SGP	0.94	0.95	0.99	0.99	1.00
Scale 3 ($k = 0.617$)					
Gain	1.00	0.61	0.60	0.61	0.58
Residual	0.61	1.00	0.94	0.96	0.94
ECPR	0.60	0.94	1.00	0.98	0.96
PRR	0.61	0.96	0.98	1.00	0.97
SGP	0.58	0.94	0.96	0.97	1.00

(1), with the different values of k and the known between-grade score correlation of about .73. Consistent with findings from other research (Castellano & Ho 2013), the correlations among the three CSPR measures were very high on all three scales. The lowest was .96 between ECPR and SGP for the manipulated scale with severe scale shrinkage.

The correlations between residual gain scores and PRRs were consistently high on all three scales. As pointed out earlier, PRRs are the percentile ranks of residual gain scores, therefore these two measures are perfectly related. The less-than-perfect correlation between residual gain scores and PRRs was due to the non linearity of their relationship. The values of the correlations were similar across all three scales. These results were consistent with expectations in Hypothesis (1). The correlations between residual gain scores and each of the CSPR measures in Table 1 were also high (.94–.96). These correlations were also similar across the three scales. These results were consistent with expectations in Hypothesis (2).

The observed correlations between simple gain scores and the CSPRs were consistent with expectations in Hypothesis (3). Due to the high correlations between residual gain scores and the CSPRs, it was expected that the correlations between simple gain scores and the CSPRs would be similar to the correlations between simple gain and residual gain scores. As shown in Table 1, the correlations between simple gain scores and the CSPR measures were slightly lower than but close to the correlations between simple gain scores and residual gain scores. When simple gain scores and residual gain scores were highly correlated, the correlations between simple gain scores and the CSPRs were also high, as demonstrated in Scale 1 and Scale 2; when the correlation between simple gain and residual gain score dropped to .61 on Scale 3, the correlations between simple gain scores and the three CSPR measures also dropped to .61 or lower.

Scatter plots of these measures on the three scales are presented in Fig. 2, in which Gain and Residual refer to simple gain scores and residual gain scores, respectively. The scatterplots provided additional information regarding the relationships between these measures beyond the correlations. First, the plots showed nonlinear relationships not only between residual gain scores and PRRs but also between each of the three CSPR measures and the simple gain or residual gain scores. The nonlinear relationships led to the under estimation of the association between these measures when using the Pearson correlation statistic. These plots also showed that compared with the gain score metrics, the percentile rank metrics enlarged the differences between individuals in the middle of the score scale where there were more examinees and reduced differences between individuals at the extremes of the scales where there were fewer examinees. Second, even when the correlations among the CSPR measures were very high (e.g., .99 on Scale 2), there were still many outliers. Further investigations revealed that these outliers were mainly examinees at the extremes of the scale. When extreme scores with few examinees were excluded, the number of outliers decreased. Third, a closer comparison of the scatter plots among the CSPR measures showed that SGPs and ECPRs were more closely related to each other than with PRRs, which could be explained by the greater conceptual similarity between the first two. The scatter

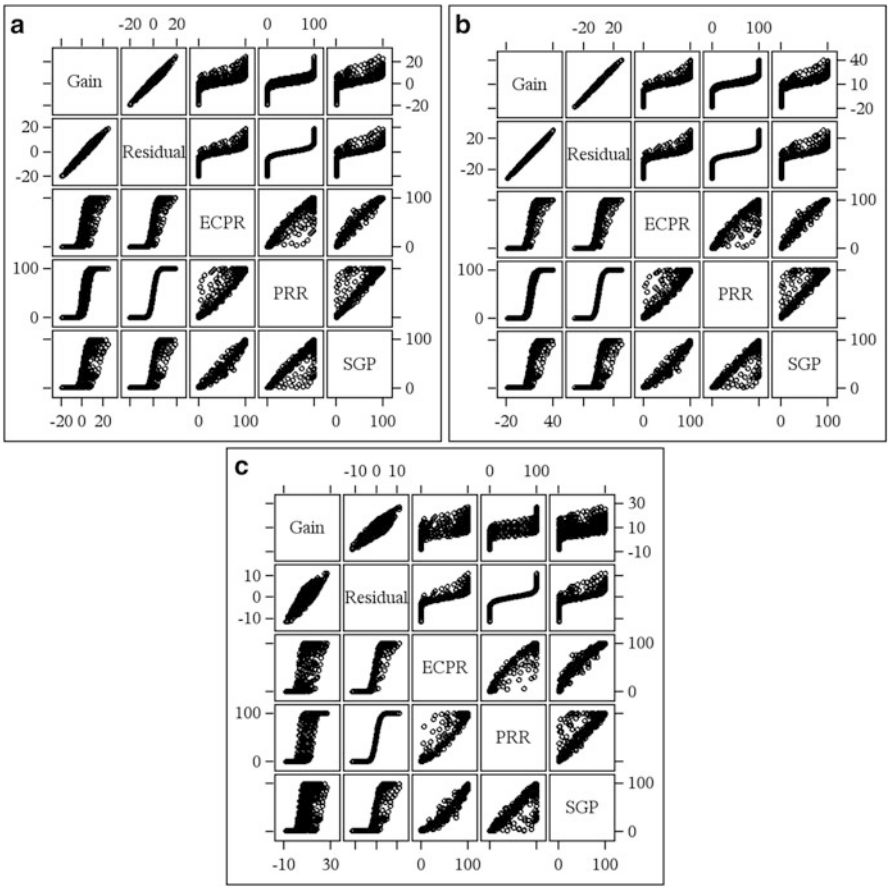


Fig. 2 Scatter plots of growth measures on (**a**) Scale 1, (**b**) Scale 2, and (**c**) Scale 3

plots also showed that for the extreme scores, students' PRRs tended to be higher than their SGPs or ECPRs on all the three scales. This observation was not expected, but might be related to the heterogeneity of the conditional score distributions.

4 Conclusion and Discussion

This study demonstrated the usefulness of a mathematical relationship as described in Li and Kolen (2011) in predicting the relationships between simple gain scores and residual gain scores on different vertical scales. Based on the conceptual similarities between some of the growth measures and findings from other research, it was hypothesized that the mathematical relationship was also useful in predicting the

approximate relationships between simple gain scores and a few CSPR measures. An empirical comparison of the growth measures was conducted which confirmed the predictions based on the mathematical relationship and other hypotheses. In addition, the empirical study also revealed some information that was not expected based on the mathematical relationship or the theoretical comparisons among the growth measures, including the many outliers that differed substantially between the highly correlated CSPR measures and the relatively higher estimates of PRRs compared to the other two CSPR measures. Though these findings might be explained by the scarcity of data for some score points or the heterogeneity of conditional score distributions, further investigations may be needed to see if the results generalize beyond the specific tests used for these analyses.

The mathematical relationship and the empirical comparison suggested that simple gain scores tended to rank order students more consistently with a few other measures, such as residual gain scores or CSPR measures, when the score variability increased from a lower grade to a higher grade, as opposed to when score variability decreased across grades. The extent to which these measures agreed also depended on the correlation of the scores between grade levels. This between-grade score correlation should remain constant among linearly related vertical scales but might differ across different grade pairs or different tests. This study was limited in that the growth measures were all based on data from two years. More research is needed to reveal the effect of vertical scales on relationships between measures based on modeling student growth trajectories over multiple years and residual gain scores or CSPRs that condition on scores from multiple prior years. Findings from this study and further investigations could provide useful information to both test developers and test users by revealing why different growth results could be highly consistent on one vertical scale but differ significantly on another vertical scale. Specifically, such information could help test developers to determine how different choices in constructing a vertical scale might impact the consistency of results from various growth measures and allow test users to understand to a greater extent how the growth measures used for decision making might have been affected by features of the particular vertical scales in use.

References

Betebenner, D. W. (2008). Toward a normative understanding of student growth. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 155–170). New York, NY: Taylor & Francis.

Betebenner, D. W. (2009). Norm-and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42–51.

Briggs, D. C., & Domingue, B. (2013). The gains from vertical scaling. *Journal of Educational and Behavioral Statistics, 38*(6), 551–576.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice, 28*(4), 3–14.

Burket, G. R. (1984). Response to Hoover. *Educational Measurement: Issues and Practice, 3*(4), 15–16.

Castellano, K. E., & Ho, A. D. (2013). Contrasting OLS and quantile regression approaches to student "growth" percentiles. *Journal of Educational and Behavioral Statistics, 38*(2), 190–215.

Dunn, J., & Allen, J. (2009). Holding schools accountable for the growth of nonproficient students: Coordinating measurement and accountability. *Educational Measurement: Issues and Practice, 28*(4), 27–41.

Goldschmidt, P., Choi K., & Beaudon, J. P. (2012). *Growth model comparison study: Practical implications of alternative models for evaluating school performance*. Retrieved from http://www.ccsso.org/Documents/2012/Growth_Model_Comparison_Study_Evaluating_School_Performance_2012.pdf.

Grady, M., Lewis, D., & Gao, F. (2010). *The effect of sample size on student growth percentiles*. Paper presented at the 2010 annual meeting of the National Council on Measurement in Education. May 1–3, Denver, CO.

Harris, D. W. (Ed.). (1963). *Problems in measuring change*. Madison: University of Wisconsin Press.

Hoover, H. D. (1984). The most appropriate scores for measuring educational development in the elementary schools: GEs. *Educational Measurement: Issues and Practice, 3*(4), 8–14.

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

Lei, P., & Zhao, Y. (2012). Effects of vertical scaling methods on linear growth estimation. *Applied Psychological Measurement, 36*(1), 21–39.

Li, D. (2015a). Investigations of the Thurstone scaling method in the ACT Aspire vertical scaling study. In W. Tao (Chair), *Constructing a vertical scale under linked scaling tests design*. Symposium conducted at the annual meeting of the National Council on Measurement in Education. April 15–19, Chicago, IL.

Li, D. (2015b). *Relationships of growth measures from different plausible vertical scales*. Paper presented at the annual meeting of the National Council on Measurement in Education. April 15–19. Chicago, IL.

Li, D., & Kolen, M. J. (2008). *Models of individual growth for school accountability—An empirical comparison*. Paper presented at the annual meeting of the American Educational Research Association. March 24–28, New York City, NY.

Li, D., & Kolen, M. J. (2011). *Relationships between status, simple gain, residual gain, and linear growth*. Paper presented at the annual meeting of the National Council on Measurement in Education. April 7–11, New Orleans, LA.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 685*, 304–305.

Maris, E. (1998). Covariance adjusted versus gain scores—Revisited. *Psychological Methods, 3*, 309–327.

Paek, I., Young, M. J., & Yi, Q. (2008). The impact of data collection design, linking method, and sample size on vertical scaling using the Rasch model. *Journal of Applied Measurement, 9*(3), 229–248.

Patz, R. (2007). *Vertical scaling in standards-based educational assessment and accountability systems*. Washington, DC: The Council of Chief State School Officers.

Phillips, S. E., & Clarizio, H. F. (1988). Conflicting growth expectations cannot both be real: A rejoinder to Yen. *Educational Measurement: Issues and Practice, 7*(4), 18–19.

Roberts, D. M., & Burrill, D. F. (1995). Gain score grading revisited. *Educational Measurement: Issues and Practice, 14*(1), 29–33.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.

Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20*(2), 227–253.

Investigation of Constraint-Weighted Item Selection Procedures in Polytomous CAT

Ya-Hui Su

Abstract To fulfill a large number of statistical and non-statistical constraints in computerized adaptive testing (CAT), the maximum priority index approaches can be used to handle many constraints simultaneously and efficiently for the construction of assessments. Many previous studies in CAT were conducted for dichotomously scored items; however, only few studies were conducted for polytomously scored items. In practice, because polytomous items are more informative, polytomous CAT tends to need fewer items than dichotomous CAT does. Many important issues in polytomous CAT need further attention. Therefore, the purpose of the study was to investigate constraint-weighted item selection procedures in polytomous CAT. The generalized partial credit model (GPCM) was considered in this study. It was found that the maximum priority index was implemented with the Fisher information, the interval information, and the posterior expected Kullback–Leibler information successfully in polytomous CAT. These three item information criteria had similar performance in terms of measurement precision, exposure control, and constraint management.

Keywords Polytomous • Constraint-weighted • Item selection • Computerized adaptive testing

1 Introduction

To construct assessments in computerized adaptive testing (CAT), the maximum priority index approaches can be used to handle a large number of statistical and non-statistical constraints simultaneously and efficiently. In previous studies, the maximum priority index approaches can be used to perform item selection not only in unidimensional CAT (Cheng & Chang 2009; Cheng, Chang, Douglas, & Guo 2009) but also in multidimensional CAT (Su 2015; Su 2016; Su & Huang 2015; Yao 2011 2012 2013). These studies on the maximum priority index approaches

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

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_7

were conducted for dichotomously scored items. In practice, Likert-type items are commonly considered in psychological inventory. Subjects specify their level of agreement or disagreement on a symmetric agree-disagree scale for a series of statements. Many previous CAT studies were conducted for dichotomously scored items; however, only a few CAT studies were conducted for polytomously scored items (Veldkamp 2003). One of the findings from these studies is that the polytomous CAT tends to need fewer items than dichotomous CAT does because polytomous items are more informative. Therefore, many issues in polytomous CAT still need further attention.

In the literature, three item information criteria have been used for polytomous CAT: maximum Fisher information (Dodd, De Ayala, & Koch 1995), maximum interval information (van Rijn, Eggn, Hemker, & Sanders 2002), and posterior expected Kullback–Leibler information (Veldkamp & van der Linden 2002). The Fisher information, which is based on ability estimate, is commonly used in CAT. In the beginning of the administration, the ability estimate is not precise to close to the true ability level. Therefore, it may not be efficient to use maximum Fisher information for item selection. It leads that items with optimal properties are selected at wrong ability levels (Lord & Novick 1968). Besides, Muraki (1993) found Fisher information function might have multi-peaks when items are calibrated with the generalized partial credit model (GPCM; Muraki 1992). In practice, an item with multi-peaks might contain more information for a small interval around the ability estimate than the item that contains maximum Fisher information at the ability estimate (van Rijn et al. 2002). To overcome these problems, van Rijn et al. (2002) proposed maximum interval information as item information criterion for polytomous CAT. The interval information function is to integrate the Fisher information function over a small interval around the ability estimate instead of maximizing Fisher information function. Chang and Ying (1996) proposed Kullback–Leibler information, a global information criterion, for item selection. In general, Kullback–Leibler information in CAT measures the distance between two likelihoods over true ability and any other ability parameter. Because the true ability of the examinee is unknown, the posterior expected information of ability is used (van der Linden 1998). The posterior expected Kullback–Leibler information is proposed for item selection (Veldkamp & van der Linden 2002).

Since the maximum priority index approaches were investigated only for dichotomously scored items, they haven't been investigated for polytomouly scored items. The maximum priority index integrating with different item information criteria might perform differently to select items for administration in polytomous CAT. In the previous studies, the GPCM is commonly considered in polytomous CAT (Veldkamp 2003; Zhou & Reckase 2014). Therefore, the purpose of the study was to investigate the performance of the constraint-weighted item selection procedures with these three item information criteria in polytomous CAT using the GPCM through simulations.

1.1 The Maximum Priority Index (MPI) Method

The maximum priority index (MPI) method can be used to monitor several statistical and non-statistical constraints simultaneously (Cheng & Chang 2009). K is the total number of constraints. $c_{ik} = 1$ represents constraint k relevant to item i and $c_{ik} = 0$ otherwise. Each constraint k is given a weight w_k to match its importance. The priority index of item i can be computed as

$$\mathrm{PI}_i = I_i \prod_{k=1}^{K} (w_k f_k)^{c_{ik}}, \tag{1}$$

where I_i is the Fisher information of item i evaluated at the current $\widehat{\theta}$. In fact, the Fisher information can be replaced with other item information criteria, such as interval information (Veerkamp & Berger 1997) or Kullback–Leibler information (Chang & Ying 1996). For a content constraint k, the priority index can be considered in a certain content area. If X_k is the number of items required from the content area, after x_k items have been selected, f_k is defined as

$$f_k = \frac{(X_k - x_k)}{X_k}. \tag{2}$$

For item exposure control constraint k, f_k can be defined as

$$f_k = \frac{1}{r_{\max}} \left(r_{\max} - \frac{n}{N} \right), \tag{3}$$

where $r_{\max}$ is the maximum item exposure rate, N is the number of examinees who have taken the CAT, and n is the number of examinees have seen item i. The item with the largest priority index will be chosen for administration.

When flexible content balancing constraints are considered, l_k and u_k are lower and upper bounds of content area k, respectively. Let μ_k is the number of items to be selected from content area k. Then,

$$l_k \leq \mu_k \leq u_k, \tag{4}$$

and

$$\sum_{k=1}^{K} \mu_k = L, \tag{5}$$

where L is test length. To incorporate both upper and lower bounds for a one-phase item selection strategy, Su and Huang (2015) suggested that f_k can be replaced with $f_{1k} f_{2k}$, which f_{1k} and f_{2k} are defined as

$$f_{1k} = \frac{1}{u_k} (u_k - x_k), \tag{6}$$

and

$$f_{2k} = \frac{(L - l_k) - (t - x_k)}{L - l_k}, \tag{7}$$

respectively. f_{1k} represents the closeness to the upper bound whereas f_{2k} represents the closeness to the lower bound. t is the number of items that have already been administered and $t = \sum_{k=1}^{K} x_k$. When f_{2k} is equal to 0, the sum of items from other constraints has reached its maximum; $f_{1k}f_{2k}$ is defined as 1 to ensure that items from constraint k can be still included for item selection. It was found the weighted mechanism successfully addresses the constraints. This method not only helps to a great extend balancing item exposure rates, but also improves measurement precision.

2 Method

2.1 Simulation Study

In this study, the GPCM (Muraki 1992) was used for data generation. For item i, the probability of obtaining a score in category v is defined as

$$P_{ik} = \frac{\exp \sum_{t=0}^{v} a_i (\theta - b_{it})}{\sum_{s=0}^{m} \exp \sum_{t=0}^{s} a_i (\theta - b_{it})}, \tag{8}$$

where a_i is the slop parameter of item i, b_{iv} is the item category parameters of item i, $v = \{0, 1, 2, \ldots, m\}$ is a category number, and θ is the ability of the examinee. Three hundred and sixty items with 6 points were generated to form a two-content pool, in which 40 % and 60 % items measured the first and the second contents, respectively. The discrimination parameters were drawn from a uniform distribution on the interval of real numbers (0.5, 1.5), difficulty parameters were drawn from a standard normal distribution, and guessing parameters were drawn from a uniform distribution on (0, 0.4). All 1000 simulated examinees were drawn from a standard normal distribution.

Four constraints were considered in the study, including content balancing, item exposure control, and item information criterion. These constraints and the corresponding weights, upper bounds, and lower bounds list in Table 1. The upper and lower bounds for the first content area were 6 and 10 items, respectively. The upper and lower bound for the second content area were 10 and 15 items,

Table 1 Constraints and weights for item selection

Constraints	Weight	Lower bound	Upper bound
Content area 1	1	6	10
Content area 2	1	10	15
Item exposure control	1		0.2
Item information criterion	1		

Note: Three item information criteria, the Fisher information (FI), the interval information (II), and the posterior expected Kullback–Leibler information (PEKLI), were considered for item selection in polytomous CAT

respectively. The maximum item exposure rates were set at .20. Three item information criteria, the Fisher information (FI), the interval information (II), and the posterior expected Kullback–Leibler information (PEKLI), were considered for item selection in polytomous CAT. When the FI criterion was considered, Fisher information function for a single item was defined as

$$I_i(\theta) = \mathrm{a}_i^2 \left[\sum_{k=1}^{m} k^2 P_{ik}(\theta) - \left(\sum_{k=1}^{m} k P_{ik}(\theta) \right)^2 \right]. \tag{9}$$

When the II criterion was considered, the interval information function (van Rijn et al. 2002) for a single item was defined as

$$\text{Interval Information Function} = \int_{\widehat{\theta}\text{-}\delta}^{\widehat{\theta}+\delta} I_i(\theta)\, d\theta, \tag{10}$$

where δ is a small constant defining the width of the interval. When the PEKLI criterion was considered, Kullback–Leibler information for a single item was defined as

$$K_i(\theta, \theta_0) \equiv \sum_{k=1}^{m} P_{ik}(\theta_0) \ln \left(\frac{P_{ik}(\theta_0)}{P_{ik}(\theta)} \right). \tag{11}$$

Because the true ability of the examinee is unknown, the posterior expected information of ability is used (van der Linden 1998). After $(t-1)$ items are administered, the PEKLI criterion (Veldkamp & van der Linden 2002) was defined as

$$KL_i\left(\widehat{\theta}^{t-1}\right) \equiv \int_{\theta} K_i\left(\theta, \widehat{\theta}^{t-1}\right) f\left(\theta \middle| u_{i1}, \ldots, u_{i_{t-1}}\right) d\theta. \tag{12}$$

During item selection in polytomous CAT, one of the three item information criteria would be used to integrate with the MPI item selection method in Eq. (1). The item

with the largest priority index in Eq. (1) would be chosen to administer. A fixed test length of 20 items was considered in this study. The expected a posteriori (EAP) estimation was used to estimate $\widehat{\theta}$.

2.2 Evaluation Criteria

The results of the simulation study were analyzed and discussed based on the following criteria: (a) measurement precision, (b) exposure control, and (c) constraint management. With respect to measurement precision, latent trait recovery was evaluated with the bias (bias) and root mean squared error of estimation (RMSE). The formulas for bias and RMSE were given as follows:

$$\text{bias} = \frac{1}{N}\sum_{n=1}^{N}\left(\widehat{\theta}_n - \theta_n\right), \tag{13}$$

and

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\widehat{\theta}_n - \theta_n\right)^2}, \tag{14}$$

where $\widehat{\theta}_n$ and θ_n are the estimated and true abilities, respectively.

With respect to exposure control, for each item information criterion, the maximum item exposure rate and the number of unused items were reported. In addition, the χ^2 statistic was used to measure the skewness of item exposure rate distribution (Chang & Ying 1999)

$$\chi^2 = \frac{1}{L/I}\sum_{i=1}^{I}(r_i - L/I)^2, \tag{15}$$

where r_i is the exposure rate of item *i*. L is the test length and I is the number of items in the pool. The smaller the χ^2 statistic, the better the item exposure control. With respect to constraint management, the number of violated constraints in each test was recorded. For each item information criterion, the averaged number of violated constraints was calculated over all examinees.

3 Results

The results of the simulation study were summarized according to measurement precision, exposure control, and constraint management.

With respect to measurement precision, the bias and RMSE of the three item information criteria list in Table 2. The PEKLI criterion obtained the best measurement precision with the smallest value in bias and RMSE. The bias of the FI, the II, and the PEKLI criteria were .012, .016, and .010, respectively. The RMSE of the FI, the II, and the PEKLI criteria were .311, .335, and .282, respectively. In general, these three item information criteria obtained very similar measurement precision. This was because the overlap rates of the selected items among three item information criteria were between 83 % and 92 %. For a 20-item test, on average the number of non-overlapping items was less or equal to four. It means applying different item information criteria would only result in four different items and suggests that there would not be many differences in measurement precision.

With respect to exposure control, the maximum item exposure rate, the number of unused items, and the χ^2 statistic of the three item information criteria list in Table 3. The three item information criteria performed similar in exposure control. The maximum item exposure rates of the three item information criteria were all smaller than 0.2. The maximum item exposure rates of the FI, the II, and the PEKLI criteria were .181, .176, and .182, respectively. The II criterion yielded the smallest value in maximum item exposure rate. The number of unused items of the FI, the II, and the PEKLI criteria were 30, 35, and 33 items, respectively. The FI criterion yielded the smallest value in the number of unused items. The χ^2 statistic of the FI, the II, and the PEKLI criteria were 9.553, 11.397, and 10.285, respectively. The FI criterion yielded the smallest χ^2 statistic. With respect to constraint management, zero averaged violations were obtained when three item information criteria were applied.

Table 2 Measurement precision when the three item information criteria were applied

Item information criteria	Bias	RMSE
Fisher information (FI)	.012	.311
Interval information (II)	.016	.335
Posterior expected Kullback–Leibler information (PEKLI)	.010	.282

Table 3 Exposure control when the three item information criteria were applied

Item information criteria	Maximum item exposure rate	Unused items	Chi-square statistic
Fisher information (FI)	.181	30	9.553
Interval information (II)	.176	35	11.397
Posterior expected Kullback–Leibler information (PEKLI)	.182	33	10.285

4 Discussion

The maximum priority index approaches can be used to handle many constraints simultaneously and efficiently for the construction of assessments in unidimensional and multidimensional CAT (Cheng et al. 2009; Cheng & Chang 2009; Su 2015; Su 2016; Su & Huang 2015; Yao 2011 2012 2013). Many previous studies in CAT focus on dichotomously scored items; however, only few studies focus on polytomously scored items (Veldkamp 2003). Because polytomous items provide more information than dichotomous items do, studies in polytomous CAT need further investigation. In this study, the MPI was implemented with the FI, the II, and the PEKLI criteria successfully in polytomous CAT using the GPCM. These three item information criteria had similar performance in terms of measurement precision, exposure control, and constraint management.

The MPI item selection method can be implemented easily and computed efficiently. The research findings from this study will advance our knowledge for item selection in polytomous CAT. However, this study has some limitations that can be addressed in future work. First, many educational and psychological tests are developed under a multidimensional framework. Items of correlated dimensions can provide information to lead to greater measurement efficiency, such as greater precision or reduced test lengths (Segall 1996; Wang & Chen 2004). In practice, item selection in multidimensional CAT is more flexible than that in unidimensional CAT. The maximum priority index approaches might be useful to be extended for multidimensional polytomous items. Second, all three item selection criteria had some unused items. When *a*-stratification design (Chang, Qian, & Ying 2001; Chang & van der Linden 2003; Chang & Ying 1999) is considered, it can obtain better measurement precision and achieve better item usage in some degree. Third, a fixed test length of 20 items was used in this study. Different measurement precisions are obtained for different ability levels and it results in a high misclassification rate, which might be costly. To achieve the same level of precision for examinees, a stopping rule of fixed-precision can be considered in CAT.

Fourth, in psychological inventory, the generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin 1996; 1998; 2000) is also commonly employed to analyze Liket-type items. The assumption of the GGUM is different from a dominance model. In the dominance model, the probability of getting a correct answer is increased with the ability level. In contrast, the GGUM is an unfolding model; that is, there exits an idea point. A higher item score is expected when a person's ability level is close to a given item on a unidimensional latent continuum. Therefore, a higher item indicates stronger levels of agreement or attraction (Andrich 1996; Roberts 1995; Roberts, Laughlin, & Wedell 1999). When a person disagrees with an attitude item because its content is either too negative or too positive relative to his/her own opinion. The GGUM is available for dichotomous or polytomous items. Many CAT studies were conducted for dominance items. However, only few CAT studies were conducted for unfolding items (Roberts, Lin, & Laughlin 2001). In practice, an attitude measurement or personality test fit better with the unfolding models than the dominance models. It is interesting to investigate item selection in CAT using unfolding models.

References

Andrich, D. (1996). A general hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling the Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology, 49*, 347–365.

Chang, H.-H., Qian, J., & Ying, Z. (2001). *a*-stratified multistage CAT with *b*-blocking. *Applied Psychological Measurement, 25*, 333–341.

Chang, H.-H., & van der Linden, W. J. (2003). Optimal stratification of item pools in alpha-stratified computerized adaptive testing. *Applied Psychological Measurement, 27*, 262–274.

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

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

Cheng, Y., & Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 62*, 369–383.

Cheng, Y., Chang, H.-H., Douglas, J., & Guo, F. (2009). Constraint-weighted *a*-stratification for computerized adaptive testing with nonstatistical constraints: Balancing measurement efficiency and exposure control. *Educational and Psychological Measurement, 69*, 35–49.

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19*, 5–22.

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 19*, 159–176.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17*, 351–363.

Roberts, J. S. (1995). Item response theory approaches to attitude measurement (Doctoral dissertation, University of South Carolina, Columbia, 1995). *Dissertation Abstracts International*, 56, 7089B.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (1996, June). *A generalized item response model for unfolding responses from a graded scale*. Paper presented at the 61st annual meeting of the Psychometric Society, Banff, Alberta, Canada.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (1998). *The generalized graded unfolding model: A general parametric item response model for unfolding graded responses* (Research Report No.RR-98-32). Princeton NJ: Educational Testing Service.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general model for unfolding unidimensional polytomous responses using item response theory. *Applied Psychological Measurement, 24*, 3–32.

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement, 59*, 211–233.

Roberts, J. S., Lin, Y., & Laughlin, J. E. (2001). Computerized adaptive testing with the generalized graded unfolding model. *Applied Psychological Measurement, 25*, 177–196.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331–354.

Su, Y.-H. (2015). The performance of the modified multidimensional priority index for item selection in variable-length MCAT. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & S.-M., Chow (Eds.), *Quantitative psychology research: The 79th annual meeting of the psychometric society* (pp. 89-97). Switzerland: Springer.

Su, Y.-H., & Huang, Y.-L. (2015). Using a modified multidimensional priority index for item selection under within-item multidimensional computerized adaptive testing. In R. E. Millsap, D. M. Bolt, L. A. van der Ark, & W.-C. Wang (Eds.), *Quantitative psychology research: The 78th annual meeting of the psychometric society* (pp. 227–242). Switzerland: Springer.

Su, Y.-H. (2016). A comparison of constrained item selection methods in multidimensional computerized adaptive testing. *Applied Psychological Measurement*. doi: 10.1177/0146621616639305

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*, 201–216.

van Rijn, P. W., Eggn, T. J. H. M., Hemker, B. T., & Sanders, P. F. (2002). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 26*, 393–411.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*, 203–226.

Veldkamp, B. P. (2003). Item selection in polytomous CAT. In H. Yanai, A. Okada, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics* (pp. 207–214). New York: Springer.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika, 67*, 575–588.

Wang, W.-C., & Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement, 28*, 295–316.

Yao, L. (2011, October). *Multidimensional CAT item selection procedures with item exposure control and content constraints*. Paper presented at the 2011 International Association of Computer Adaptive Testing IACAT Conference, Pacific Grove, CA.

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: Theory and applications. *Psychometrika, 77*, 495–523.

Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. *Applied Psychological Measurement, 37*, 3–23.

Zhou, X., & Reckase, M. D. (2014). Optimal item pool design for computerized adaptive tests with polytomous items using GPCM. *Psychological Test and Assessment Modeling, 56*(3), 255–274.

Estimating Classification Accuracy and Consistency Indices for Multidimensional Latent Ability

Wenyi Wang, Lihong Song, Shuliang Ding, and Yaru Meng

Abstract For criterion-referenced tests, classification consistency and accuracy are viewed as important indicators for evaluating reliability and validity of classification results. Numerous procedures have been proposed in the framework of unidimensional item response theory (UIRT) to estimate these indices. Some of these were based on total sum scores, others on latent trait estimates. However, there exist very few attempts to develop them in the framework of multidimensional item response theory (MIRT). Based on previous studies, the aim of this study is first to estimate the consistency and accuracy indices of multidimensional ability estimates from a single administration of a criterion-referenced test. We also examined how Monte Carlo sample size, sample size, test length, and the correlation between the different abilities affect the estimate quality. Comparative analysis of simulation results indicated that the new indices are very desirable to evaluate test-retest consistency and correct classification rate of different decision rules.

Keywords Multidimensional item response theory • Classification accuracy • Classification consistency • Complex decision rule

W. Wang • S. Ding
College of Computer Information Engineering, Jiangxi Normal University, 99 Ziyang Road, Nanchang, Jiangxi, P. R. China
e-mail: wenyiwang2009@gmail.com; ding06026@163.com

L. Song (✉)
Elementary Educational College, Jiangxi Normal University, 99 Ziyang Road, Nanchang, Jiangxi, P. R. China
e-mail: viviansong1981@163.com

Y. Meng
School of Foreign Studies, Xi'an Jiaotong University, 28 West Xianning Road, Xi'an, Shanxi, P. R. China
e-mail: yarum@163.com

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_8

1 Introduction

Criterion-referenced tests are the most widely used type of test in education because its goal is to determine whether teachers and schools are effectively teaching students what they are expected to learn. An important measurement practice in this context is to categorize each student into one of two, three, or more achievement levels based on a set of standards or cutoff scores. Therefore, making better decision about student achievement is the primary concern. But it is well recognized that no test score is without error, and important decisions are best made using several scores or sources of information to minimize errors.

For criterion-referenced tests, classification consistency and accuracy are important indicators to evaluate the reliability and validity of classification results. Numerous procedures have been proposed to estimate these indices in the framework of unidimensional item response theory (UIRT) (Huynh 1990; Lathrop & Cheng 2013; Lee 2010; Rudner 2001, 2005; Schulz, Kolen & Nicewander 1999; Wang, Kolen & Harris 2000; Wyse & Hao 2012). Some of these were based on total sum scores, while others on latent trait estimates (Lathrop & Cheng 2013). The Lee approach (Lee 2010) belongs to the former, whereas the Rudner approach (Rudner 2001, 2005) and its extension, the Guo approach (Guo 2006), falls into the latter category. Because the original Rudner index and Guo index are used to estimate only classification accuracy alone, Wyse and Hao (2012) then proposed the Rudner- and Guo-based classification consistency indices.

The well-known examples of criterion-referenced educational and psychological tests include American National Assessment of Educational Progress (NAEP), Program for International Student Assessment (PISA), IEA's Trends in International Mathematics and Science Study (TIMSS), Chinese National Assessment of Educational Quality (NAEQ) and personality assessments like NEO-PI-R. They are all standardized tests administered to large numbers of students and most are multidimensional to some degree (Bolt & Lall 2003; Debeer, Buchholz, Hartig & Janssen 2014; Makransky, Mortensen & Glas 2012; Rijmen, Jeon, von Davier & Rabe-Hesketh 2014; Yao & Boughton 2007; Zhang 2012). Multidimensional item response theory (MIRT) has been devoted to models that include more than one latent trait to account for the multidimensional nature of the complex constructs. For example, the overall construct of mathematics in TIMSS is defined to encompass four content domains: number, algebra, geometry, and data and chance. Up to now, it has been successfully employed to analyze these large-scale assessments.

Although MIRT has enjoyed tremendous growth, solutions to some problems remain unavailable. One case in point is the estimate of classification accuracy and consistency indices. Yao (2013) and LaFond (2014) focused on accuracy and consistency indices under MIRT based total sum scores only. It is problematic because of two reasons:for one thing, classifications made with the latent ability estimates shall be equally or more accurate than classifications made with total sum score (Lathrop & Cheng 2013), at least for graded response model; and for another,

it may be difficult to estimate accuracy and consistency indices in each content areas when some items may measure more than two domains (complex structure) or when domains score estimates are highly unreliable (Pommerich & Nicewander 1999). The current study addresses this issue based on the assumption that the cut scores are aligned either on the latent trait scale or on the total sum scores. It aims to incorporate the previous UIRT indices results into the case of MIRT and obtain the desired goal from a single administration of a test. The rest of this article proceeds as follows. Section 2 starts with a review on a MIRT model and consistency and accuracy indices for total sum scores. Section 3 introduces decision rule and Guo-based accuracy and consistency indices. Section 4 provides a simulation study and explains the simulation results. Finally, Sect. 5 presents the conclusion.

2 Model and Methods

2.1 A Multidimensional Graded Response Model

A multidimensional graded response model (MGRM) is a generalization of the unidimensional graded response model and it uses response function that has the logistic function. The parameterization of this model given here considers the lowest score on item j to be 0 and the highest score to be K_j. Let $\boldsymbol{\theta} = (\theta_1, \theta_d, \ldots, \theta_d)'$ denote a set of d latent abilities, $g(\boldsymbol{\theta})$ denote the distribution of ability, $\boldsymbol{\alpha}_j$ be a vector of discrimination parameters on item j, and β_{jk} be a threshold parameter related to item difficulty with which a person will reach the kth step of item j. Given an examinee with the multidimensional ability vector $\boldsymbol{\theta}$, his probability of successfully performing the work in step k or more advanced steps in answering item j can be written as:

$$P\left(y_{ij} \geq k \middle| \boldsymbol{\theta}, \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j\right) = \frac{1}{1 + exp\left(\beta_{jk} - \boldsymbol{\alpha}_j'\boldsymbol{\theta}\right)}, \tag{1}$$

where $j = 1, 2, \ldots, J$ and $k = 0, 1, \ldots, K_j + 1$.

The probability of receiving a specific score k is the difference between the probability of successfully performing the work for step k or more advanced steps and that of the work for $k + 1$ or more advanced steps. Then the probability that an examinee will receive a score of k is

$$\begin{aligned} P_{jk}(\boldsymbol{\theta}) &= P\left(y_j = k \middle| \boldsymbol{\theta}, \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j\right) \\ &= P\left(y_{ij} \geq k \middle| \boldsymbol{\theta}, \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j\right) - P\left(y_j \geq k + 1 \middle| \boldsymbol{\theta}, \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j\right). \end{aligned} \tag{2}$$

A number of computer programs have been developed for estimating item and ability parameters in MGRM, such as BMIRT (Yao 2003), IRTPRO program

(Cai, Thissen & du Toit 2011), MIRT package for the R Environment (Chalmers 2012; R Core Team, 2015) and so on. Before we consider consistency and accuracy indices, we will first present some formulas. Suppose that item and ability parameters are estimated, the marginal probability of the total summed score X is given by

$$P(X=x)=\int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{+\infty}\int P_J\left(X=x\middle|\boldsymbol{\theta}\right)g(\boldsymbol{\theta})\,d\theta_1\ldots d\theta_d, \tag{3}$$

where $x=\sum_{j=1}^{J}y_j$ is a particular realization of the total summed score for a student and $P\left(X=x\middle|\boldsymbol{\theta}\right)$ is the conditional distribution of X. Due to the IRT's assumption of conditional independence of the responses given the $\boldsymbol{\theta}$-vector, the conditional probability of the summed score x can be written as follows:

$$P_J\left(X=x\middle|\boldsymbol{\theta}\right)=\sum_{k=0}^{K_j}P_{J-1}\left(X=x-k\middle|\boldsymbol{\theta}\right)P_{Jk}(\boldsymbol{\theta}), \tag{4}$$

where $P_{Jk}(\boldsymbol{\theta})$ is defined by Eq. (2).

Assuming conditional independence of the responses given the $\boldsymbol{\theta}$-vector, the likelihood function of the observed data $\mathbf{y}_i$ is

$$L\left(\mathbf{y}_i\middle|\boldsymbol{\theta}_i,\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j\right)=\prod_{j=1}^{J}\prod_{k=0}^{K_j}P\left(y_{ij}=k\middle|\boldsymbol{\theta}_i,\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j\right)^{1_{(y_{ij}=k)}}, \tag{5}$$

where an indicator function is defined as

$$1_{(y_{ij}=k)}=\begin{cases}1 & \text{if } y_{ij}=k\\ 0 & otherwise\end{cases}. \tag{6}$$

2.2 *Consistency Indices for Summed Scores Using MIRT model*

First, we briefly describe classification consistency indices for summed scores using MIRT model (Yao 2013), which are based on the Lee approach (Lee 2010). Let us assume now that the test scores on one test form are used to classify students in c categories defined by c + 1cutoff scores $s_0, s_1, \ldots, s_c$, with $0 = s_0 < s_1 < \cdots < s_{c-1} < s_c = +\infty$ and $s_{c-1} < \sum_j^J K_j$. That is, examinees with an observed score less than s_1 are assigned to the first category; examinees with a score greater than or equal to s_1 and less than s_2 are assigned to the second category. Finally, in the same

manner, examinees with a score greater than or equal to s_{c-1} are assigned to the cth category.

Given the conditional distribution of X and the cut scores, the conditional probability of scoring in each performance category can be computed by summing up the conditional probabilities of all total summed-score x values that belong to category h, as follows:

$$p_{\boldsymbol{\theta}}(h) = \sum_{\{x: s_{(h-1)} \leq x < s_h\}} P_J(X = x|\boldsymbol{\theta}), \tag{7}$$

where $h = 1, 2, \ldots, c$.

The conditional classification consistency index $\phi(\boldsymbol{\theta})$ is defined as the probability that an examinee with $\boldsymbol{\theta}$ is classified into the same category in two independent administrations of parallel forms of a test, and it can be written as

$$\phi\,(\boldsymbol{\theta}) = \sum_{h=1}^{c} [p_{\boldsymbol{\theta}}(h)]^2 \tag{8}$$

The conditional classification consistency index quantifies classification consistency for different levels of θ. The marginal classification consistency index ϕ is given by

$$\phi = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \phi\,(\boldsymbol{\theta})\, g\,(\boldsymbol{\theta})\, d\theta_1 \ldots d\theta_d. \tag{9}$$

Another well-known index, a κ coefficient is computed as

$$\kappa = \frac{\phi - \phi_c}{1 - \phi_c}. \tag{10}$$

where ϕ_c is a chance probability. As typically defined (Cohen 1960; Huynh 1976), the chance probability is computed by $\phi_c = \sum_{h=1}^{H} [p(h)]^2$, where p(h) corresponding to the marginal category probability obtained by integrating the conditional category probabilities over the θ distribution and can be written as

$$p(h) = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} p_{\boldsymbol{\theta}}(h) g\,(\boldsymbol{\theta})\, d\theta_1 \ldots d\theta_d. \tag{11}$$

2.3 Accuracy Indices for Summed Scores Using MIRT model

Now, suppose we have a set of true cut scores on the summed-score metric, $\tau_0 = 0$, $\tau_1, \ldots, \tau_H$, we need to determine the true category of each examinee with $\boldsymbol{\theta}$ or

τ (i.e., expected summed score). The expected summed score for a student with ability $\boldsymbol{\theta}$ is obtained by

$$\tau(\boldsymbol{\theta}) = \sum_{j=1}^{J} \sum_{k=0}^{K_j} k P_{jk}(\boldsymbol{\theta}). \tag{12}$$

Then by comparing τ with the true cut scores, we know the classification τ for this examinee. If τ is classified into the hth category, that is, $\tau \in [\tau_h, \tau_{h+1})$, then the hth category is assumed as the "true" category of this examinee. The conditional classification accuracy index $\gamma(\boldsymbol{\theta})$ is defined as the probability that an examinee with $\boldsymbol{\theta}$ is classified into the "true" category assuming known cut scores on a single test, and it can be written as

$$\gamma(\boldsymbol{\theta}) = p_{\boldsymbol{\theta}}(h), \textit{ for } \tau(\boldsymbol{\theta}) \in [\tau_h, \tau_{h+1}). \tag{13}$$

The marginal classification accuracy index, γ, is given by

$$\gamma = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \gamma(\boldsymbol{\theta})\, g(\boldsymbol{\theta})\, d\theta_1 \ldots d\theta_d. \tag{14}$$

The conditional false positive error rate is defined here as the probability that an examinee is classified into a category that is higher than the examinee's true category, which is expressed as

$$\gamma^{+}(\boldsymbol{\theta}) = \sum_{h'=h+1}^{H} p_{\boldsymbol{\theta}}\left(h'\right), \textit{ for } \tau(\boldsymbol{\theta}) \in [\tau_h, \tau_{h+1}). \tag{15}$$

By contrast, the conditional false negative error rate is the probability that an examinee is classified into a category that is lower than his true category, which is given by

$$\gamma^{-}(\boldsymbol{\theta}) = \sum_{h'=1}^{h-1} p_{\boldsymbol{\theta}}\left(h'\right), \textit{ for } \tau(\boldsymbol{\theta}) \in [\tau_h, \tau_{h+1}). \tag{16}$$

The marginal false positive and false negative error rates of $\gamma^{+}$and γ^{-} are

$$\gamma^{+} = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \gamma^{+}(\boldsymbol{\theta})\, g(\boldsymbol{\theta})\, d\theta_1 \ldots d\theta_d. \tag{17}$$

and

$$\gamma^{-} = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \gamma^{-}(\boldsymbol{\theta})\, g(\boldsymbol{\theta})\, d\theta_1 \ldots d\theta_d. \tag{18}$$

3 Decision Rule, Consistency Index and Accuracy Index

3.1 Decision Rule for Multidimensional Latent Ability

Complex decision rules may be motivated by the desire to increase the reliability, accuracy, or validity of the resulting decision (Douglas & Mislevy 2010). For more discussion of decision rules, please refer to, for example, Douglas and Mislevy (2010), who presented the decision rules based on multiple scores in great details. For example, you can be qualified for admission as a postgraduate if you have achieved a minimum score on each test and earned the required total score in postgraduate admission examination.

For the above example, without the loss of generality, we can assume that you need to estimate accuracy index of the kth dimension for multiple classifications, this is achieved via the following decision regions:

$$R_{1k} = \left\{(\theta_1, \theta_2, \ldots, \theta_d) \,\middle|\, \tau_{0k} \le \theta_k < \tau_{1k}, -\infty < \theta_{k'} < +\infty, k' = 2, 3, \ldots, d\right\} \quad (19)$$

$$R_{hk} = \left\{ (\theta_1, \theta_2, \ldots, \theta_d) \,\middle|\, \tau_{(h-1)k} \le \theta_k < \tau_{hk}, -\infty < \theta_{k'} < +\infty, k' = 1, \ldots, k-1, k+1, \ldots, d\right\} \quad (20)$$

where $h = 2, 3, \ldots, H$, $\tau_{(h-1)k}$ is a cut score on the kth dimension, and $-\infty = \tau_{0k} < \tau_{1k} < \cdots < \tau_{Hk} = +\infty$.

If you need to estimate accuracy index of overall score for multiple classifications, this is achieved via the following decision regions:

$$R_{1(H+1)} = \left\{(\theta_1, \theta_2, \ldots, \theta_d) \,\middle|\, \tau_{0(H+1)0} < \sum\nolimits_{k=1}^{d} w_k\theta_k < \tau_{1(H+1)}\right\} \quad (21)$$

$$R_{h(H+1)} = \left\{(\theta_1, \theta_2, \ldots, \theta_d) \,\middle|\, \tau_{(h-1)(H+1)} \le \sum\nolimits_{k=1}^{d} w_k\theta_k < \tau_{h(H+1)}\right\} \quad (22)$$

where $h = 2, 3, \ldots, H$, w_k is a weight on the kth dimension, $\tau_{(h-1)(H+1)}$ is a cut score on overall score and $-\infty = \tau_{0(H+1)} < \tau_{1(H+1)} < \cdots < \tau_{H(H+1)} = +\infty$.

Furthermore, when the decision rule incorporates each of domains and overall score rather than a single domain or overall score, the conjunctive–compensatory rule was structured to require a minimum score on each domain and a more stringent overall weighted score. Decision regions is defined on the metric of $\boldsymbol{\theta}$ scale scores as follows

$$R_h = \left\{(\theta_1, \theta_2, \ldots, \theta_d) \,\middle|\, \tau_{(h-1)k} \le \theta_k, k = 1, 2, \ldots, d, \tau_{(h-1)(H+1)} \le \sum\nolimits_{k=1}^{d} w_k\theta_k\right\} - \cup_{h'=h+1}^{H} R_{h'}, \quad (23)$$

where $\tau_{(h-1)k}$ is a cut score on the kth dimension, $\tau_{(h-1)(H+1)}$ is a cut score on overall score, w_k is a weight on the kth dimension, and $h = 1, 2, \ldots, H$.

3.2 Guo-Based Accuracy and Consistency Indices

In this section, we extended the Guo approach to estimate consistency and accuracy indices for multidimensional latent ability. This approach is computationally easy and can be directly adapted into MIRT because its UIRT framework is closely tied to the normalized likelihood.

Guo (2006) defined classification accuracy index as the percentage of agreement between the observed and expected proportions of examinees in each of the categories under the UIRT framework. In the framework of MIRT, consistency and accuracy indices are required to estimate each dimension or domain score and overall score. In the following, we will expand this approach which is suitable for estimating consistency and accuracy indices for complex decision rules in MIRT into the new context. Our approach can be used to estimate consistency and accuracy indices for each domain and overall score. Suppose the $\boldsymbol{\theta}$ space can be partitioned into H separate decision regions, $R_1, \ldots, R_{\mathrm{H}}$, corresponding to the various categories, we can determine the true category of each examinee with $\boldsymbol{\theta}$. In other words, a decision rule is a function from the $\boldsymbol{\theta}$ space into the set of categories. From the idea of the Guo approach, the expected probability of scoring in any particular category can be obtained using the likelihood functions as

$$p_{ih} = p_i(R_h) = \frac{\int_{R_h} L\left(\mathbf{y}_i \middle| \boldsymbol{\theta}, \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j\right) d\boldsymbol{\theta}}{\sum_{h=1}^{H} \int_{R_h} L\left(\mathbf{y}_i \middle| \boldsymbol{\theta}, \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j\right) d\boldsymbol{\theta}}, \tag{24}$$

where $L\left(\mathbf{y}_i \middle| \boldsymbol{\theta}, \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j\right)$ is defined by Eq. (5) and h = 1,2, . . . ,H.

For a test data set with a particular sample size N, given an already calibrated set of item parameters, the ability vector could be estimated either via (weighted) maximum likelihood estimation (MLE) or using Bayesian methods such as maximum a posteriori (MAP) estimation or expected a posteriori (EAP) estimation (Wang 2015). Let a N by H matrix of weights $\mathbf{W}$ denote the flag of the performance-level category that the examinee obtained on the test. The weight matrix $\mathbf{W}$ can be formulated based on the examinee's ability estimate and by comparing those estimates with the cut-scores. That is, the entry w_{ih} in row i and column h of $\mathbf{W}$ is 1 if the examinee's ability estimate is classified into performance level category h, and 0 otherwise. Then a classification accuracy index can be written as

$$\gamma = \frac{\sum_{i=1}^{N} \sum_{h=1}^{H} (p_{ih} * w_{ih})}{N}. \tag{25}$$

The κ coefficient related to the Guo-based consistency index is computed as

$$\kappa = \frac{\gamma - \gamma_c}{1 - \gamma_c}. \tag{26}$$

where $\gamma_c = \sum_{h=1}^{H} (p_{\bullet h} * w_{\bullet h}) = \sum_{h=1}^{H} \left(\sum_i p_{ih}/N\right) \left(\sum_i w_{ih}/N\right)$.

Similar to the Eqs. (17) and (18), the marginal false positive and false negative error rates of γ^+ and γ^-, respectively, are

$$\gamma^+ = \frac{\sum_{i=1}^{N} \sum_{h=1}^{h_i - 1} p_{ih}}{N}, \tag{27}$$

and

$$\gamma^- = \frac{\sum_{i=1}^{N} \sum_{h=h_i+1}^{H} p_{ih}}{N}, \tag{28}$$

where $h_i = \arg\max_h (w_{ih})$ represents the observed performance level category of the examinee i.

By comparison, classification consistency provides a measure of the proportion of students who would be classified into the same category on parallel replications of the same test. A classification consistency index can be expressed as

$$\phi = \frac{\sum_i \sum_h (p_{ih} * p_{ih})}{N}. \tag{29}$$

The κ coefficient related to the Guo-based consistency index is computed as

$$\kappa = \frac{\phi - \phi_c}{1 - \phi_c}. \tag{30}$$

where $\phi_c = \sum_{h=1}^{H} (p_{\bullet h})^2 = \sum_{h=1}^{H} \left(\sum_i p_{ih}/N\right)^2$.

4 Simulation Study

Given that the classification consistency and accuracy indices based on the Guo approach is new to multidimensional latent ability, an important question is whether the Guo-based indices can accurately estimate true values of simulation indices. The true accuracy is the proportion of times that each examinee is classified into the true ability category by comparing the observed performance with the known

cut-scores. The true consistency is the proportion of times that each examinee is classified into the same category on two parallel tests. This is also called estimated test–retest consistency rate.

4.1 Design of Experiment

A simulation study following the MGRM is conducted. Several factors were manipulated in this study. The dimensions were initialized to 1, 2, and 4 respectively. In a two- or four-factor model, three levels of correlation between pairs of dimensions, $\rho = .00$, $\rho = .50$, and $\rho = .80$ were considered. Sample size could impact accuracy of parameter estimates, and in turn the classification indices. The sample consisted of either $N = 1000$ or $N = 3000$ examinees. A sample size of $N = 1000$ was chosen as the lower bound (Yao & Boughton 2007). The ability vectors were generated from multivariate normal distributions with an appropriately sized mean vector of 0 and covariance matrix Σ, where the diagonal elements of Σ were all 1 and the off-diagonal elements were given by the correlation for the associated condition.

It is well accepted that a good validity or reliability index should be sensitive to changes in test length. Doubling the number items (or increasing the test to two times the length) may improve the reliability of a short test substantially. Therefore, the accepted notion of test length was as follows: test length for the one-factor model is either 10 or 20; for a two-factor model is either 15 or 30 and for a four-factor model is either 30 or 60. In order to balance the information of the domains or dimensions (Yao 2012, 2014), content balancing techniques can be used to ensure that tests fulfill requirements with respect to content or domain areas, such as the number of items from various dimensions (Kroehne, Goldhammer & Partchev 2014; Yao 2012, 2014). Thus, it is necessary to impose serious constraints on the number of each dimension measured by a test. In a two-factor model, the constraints for a 15-item test are such that two five-items sets each loaded exclusively on one of the two dimensions and the remaining 5 items loaded on both of the two dimensions. Totally, there are 10 items measuring each dimension. The above simulation conditions have been often used in the literature (C. Wang 2015; C. Wang & Nydick 2015; Yao & Boughton 2007). The fully crossed design yielded a total of 28 conditions for each sample size, where each condition was replicated 10 times to estimate test–retest consistency.

Item parameters were fixed across all replications. They were originally described and used by Cai (2010) (Table 1 in Sec. 2.1) with 2 dimensions and 10 three-category items. These generated parameters were chosen to resemble values encountered in real data analysis. Considering content constraints, we used these item parameters to simulate six tests. Also, the same cut-points (50 % of maximum score, 80 % of maximum score, and a three-category classification with both 50 % and 80 %) were considered in the simulation study. For example, when a test had ten items and each item was scored against the three ordered categories, the two cut scores of 10 and 16 were used to classify examinees into one of the

Table 1 Indices estimates error

Item parameters	Monte Carlo sample size	*bias*		*abs*		*RMSE*	
		Lee	Guo	Lee	Guo	Lee	Guo
Simulation	1000	0.0035	−0.0029	0.0071	0.0110	0.0092	0.0137
	3000	0.0035	0.0027	0.0071	0.0082	0.0092	0.0102
	9000	0.0035	0.0048	0.0071	0.0081	0.0092	0.0100
Estimation	1000	0.0036	−0.0005	0.0072	0.0096	0.0093	0.0120
	3000	0.0036	0.0042	0.0071	0.0090	0.0092	0.0111
	9000	0.0036	0.0062	0.0072	0.0088	0.0092	0.0109

three categories. Note that Monte Carlo method can be used to tackle intractable summations or high-dimensional integrals. Therefore, Monte Carlo method was employed to estimate Eqs. (4) and (24). This study also examined the effect of the Monte Carlo sample size.

4.2 Results

Due to space constraints, we only present the results from the decision rule made based on total sum scores. Table 1 shows the bias, absolute bias (abs), and root mean squared error (RMSE) of two classification accuracy indices conditional on the simulating or estimating item parameters and Monte Carlo sample size across other conditions. The results suggested that: (a) the error of accuracy indices estimates tends to be smaller when the simulating item parameters are used; (b) Monte Carlo sample size has a large impact on the Guo-based index. This is because the sample space of latent abilities is not countable; and (c) the RMSE of Guo-based index estimates decrease as the Monte Carlo sample size increases. In terms of precision, the recommended Monte Carlo sample size for Lee-based index or Guo-based index is about 3000, as is shown in the following table.

Table 2 shows the simulation and estimation accuracy indices. The results indicated that: (a) on the whole, the values of two accuracy indices are very close to the corresponding true accuracy rates across all conditions; (b) as the sample size increases, the values of the accuracy indices increase accordingly in many cases; (c) the accuracy also tends to increase as the correlation between latent abilities increases; and (d) the values of the Guo-based index are equal to or higher than that of the Lee-based index for the unidimensional model. The simulation and estimation consistency indices with similar trend are not presented here.

Table 2 The simulation and estimation accuracy indices

				Simulation		Estimation		Kappa	
Dimension	Correlation	Test length	Sample size	Lee	Guo	Lee	Guo	Lee	Guo
1	NA	10	1000	0.8217	0.8278	0.8261	0.8360	0.7087	0.7219
			3000	0.8132	0.8214	0.8251	0.8329	0.6989	0.7231
		20	1000	0.8731	0.8808	0.8761	0.8824	0.7951	0.7973
			3000	0.8665	0.8719	0.8710	0.8779	0.7773	0.7782
4	0.0	30	1000	0.8846	0.8816	0.8783	0.8720	0.7675	0.7539
			3000	0.8758	0.8747	0.8758	0.8709	0.7520	0.7407
		60	1000	0.9102	0.9170	0.9155	0.9067	0.8255	0.7913
			3000	0.9145	0.9138	0.9136	0.9054	0.8329	0.8185
	0.5	30	1000	0.8873	0.8804	0.8924	0.8929	0.8139	0.8217
			3000	0.8927	0.8872	0.8942	0.8928	0.8123	0.8095
		60	1000	0.9232	0.9190	0.9258	0.9206	0.8754	0.8666
			3000	0.9306	0.9272	0.9279	0.9246	0.8705	0.8662
	0.8	30	1000	0.9096	0.9022	0.9069	0.9102	0.8363	0.8391
			3000	0.9043	0.9020	0.9071	0.9079	0.8417	0.8435
		60	1000	0.9316	0.9334	0.9341	0.9316	0.8936	0.8945
			3000	0.9339	0.9334	0.9326	0.9326	0.8828	0.8863
M				0.8920	0.8921	0.8939	0.8936	0.8115	0.8095

Note: The results of two-factor model was not shown here

5 Discussions

Successfully integrating a MIRT model as a diagnostic model into proficiency tests can better direct feedback about student strengths and weakness to significantly improving the future instruction and learning (Chang 2012). Based on previous studies (Guo 2006; Lathrop & Cheng 2013; Wyse & Hao 2012; Yao 2013), Guo-based consistency and accuracy indices have been adapted to MIRT models, and their performance with the MGRM was evaluated through simulation study. The simulation results show that Guo-based indices work well because their values match closely with test-retest consistency rates or true accuracy rates. Following the previous developments (Lathrop & Cheng 2013), this article also investigated differences in simulated classification accuracy and consistency between using latent traits estimate and total score to make classification decisions. It is shown that when data follows MGRM, it is preferable to make one's decision from the latent traits rather than the total score, and when the correlation between latent abilities is high, they were more consistent with UIRT models, such as GRM, 2PL, or 3PL model (Lathrop & Cheng 2013).

Finally, the applications and directions based on the current research are suggested. First, the potential application of these indices is to solve a practical problem in test development. Once the accuracy and consistency indices of each content or domain is provided in a pilot test with the population of interest, items measuring

specific content or domain with low indices can be created, thereby providing test developers with a way to control the specific domain contents measured by the test. Second, they might be useful in developing an item selection algorithm in adaptive tests. According to Chang (2015), to make multidimensional classification decisions, item selection is the most important procedure in adaptive testing. Third, we have only focused on point estimates of the indices. To more precisely describe these estimates, the construction of their confidence intervals for comparing different samples or different tests needs further investigation. Finally, the current paper focused only on the MGRM. In the future, these indices should be applied to many other MIRT models. This important area merits further investigation. It is also convenient to estimate consistency and accuracy indices for each dimension based on the Rudner approach (Rudner 2001, 2005) where the true cut scores are needed to set on the θ metric, or it can be determined from translating the cutscore on the summed-score metric to the ability cut-score.

Acknowledgments This research is supported by the China Scholarship Council (CSC No. 201509470001), the National Natural Science Foundation of China (Grant No. 31500909, 31360237, and 31160203), the Key Project of National Education Science "Twelfth Five Year Plan" of Ministry of Education of China (Grant No. DHA150285), the Humanities and Social Sciences Research Foundation of Ministry of Education of China (Grant No. 13YJC880060 and 12YJA740057), the National Natural Science Foundation of Jiangxi Province (Grant No. 20161BAB212044), Jiangxi Education Science Foundation (Grant No. 13YB032), the Science and Technology Research Foundation of Education Department of Jiangxi Province (GJJ13207), and the Youth Growth Fund and the Doctoral Starting up Foundation of Jiangxi Normal University. The authors thank the editor Jeffrey A. Douglas for his helpful comments and suggestions. Thanks to Prof. Hua-Hua Chang for his kindly help.

References

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Psychological Measurement, 27*(6), 395–414.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]*. Lincolnwood, IL: Scientific Software International.

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

Chang, H.-H. (2012). Making computerized adaptive testing diagnostic tools for schools. In R. W. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 195–226). Charlotte, NC: Information Age.

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika, 80*(1), 1–20.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics, 39*(6), 502–523.

Douglas, K. M., & Mislevy, R. J. (2010). Estimating classification accuracy for complex decision rules based on multiple scores. *Journal of Educational and Behavioral Statistics, 35*(3), 280–306.

Guo, F. (2006). Expected classification accuracy using the latent distribution. *Practical Assessment, Research & Evaluation, 11*(6), 1–6.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13*(4), 253–264.

Huynh, H. (1990). Computation and statistical inference for decision consistency indexes based on the Rasch model. *Journal of Educational Statistics, 15*, 353–368.

Kroehne, U., Goldhammer, F., & Partchev, I. (2014). Constrained multidimensional adaptive testing without intermixing items from different dimensions. *Psychological Test and Assessment Modeling, 56*(4), 348–367.

LaFond, L. J. (2014). *Decision consistency and accuracy indices for the bifactor and testlet response theory models Detecting Heterogeneity in Logistic Regression Models*. Unpublished doctoral dissertation, University of Iowa.

Lathrop, Q. N., & Cheng, Y. (2013). Two approaches to estimation of classification accuracy rate under item response theory. *Applied Psychological Measurement, 37*(3), 226–241.

Lee, W.-C. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement, 47*(1), 1–17.

Makransky, G., Mortensen, E. L., & Glas, C. A. W. (2012). Improving personality facet scores with multidimensional computer adaptive testing: An illustration with the Neo Pi-R. *Assessment, 20*(1), 3–13.

Pommerich, M., & Nicewander, W. A. (1999). Estimating average domain scores. *Journal of Educational Measurement, 36*(3), 199–216.

R Core Team. (2015). *R: A language and environment for statistical computing (Version 3.2)*. Vienna, Austria: R Foundation for Statistical Computing.

Rijmen, F., Jeon, M., von Davier, M., & Rabe-Hesketh, S. (2014). A third-order item response theory model for modeling the effects of domains and subdomains in large-scale educational assessment surveys. *Journal of Educational and Behavioral Statistics, 39*(4), 235–256.

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation, 7*(14), 1–8.

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment Research & Evaluation, 10*(13), 1–4.

Schulz, E. M., Kolen, M. J., & Nicewander, W. A. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement, 23*(4), 347–362.

Wang, C. (2015). On latent trait estimation in multidimensional compensatory item response models. *Psychometrika, 80*(2), 428–449.

Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. *Journal of Educational Measurement, 37*(2), 141–162.

Wang, C., & Nydick, S. (2015). Comparing two algorithms for calibrating the restricted noncompensatory multidimensional IRT model. *Applied Psychological Measurement, 39*(2), 119–134.

Wyse, A. E., & Hao, S. (2012). An evaluation of item response theory classification accuracy and consistency indices. *Applied Psychological Measurement, 36*(7), 602–624.

Yao, L. (2003). *BMIRT: Bayesian multivariate item response theory [Computer software]*. Monterey, CA: CTB/McGraw-Hill.

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: Theory and applications. *Psychometrika, 77*(3), 495–523.

Yao, L. (2013). *Classification accuracy and consistency indices for summed scores enhanced using MIRT for test of mixed item types*. Retrieved March 1, 2015, from http://www.bmirt.com/8220.html.

Yao, L. (2014). Multidimensional CAT item selection methods for domain scores and composite scores with item exposure control and content constraints. *Journal of Educational Measurement, 51*(1), 18–38.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*(2), 83–105.

Zhang, J. (2012). Calibration of response data using MIRT models with simple and mixed structures. *Applied Psychological Measurement, 36*(5), 375–398.

Item Response Theory Models for Person Dependence in Paired Samples

Kuan-Yu Jin and Wen-Chung Wang

Abstract When paired samples (e.g., spouses, couples, twins, or siblings) are surveyed, the item responses of the paired samples may be locally dependent between the paired persons because the investigated latent traits often involve a relationship between the paired persons (e.g., marriage satisfaction). Standard item response theory (IRT) models fail to consider the dependence between paired persons and thus are not applicable. In this study, we developed new IRT models to account for person dependence and conducted simulation studies to evaluate the parameter recovery of the new models and the consequences for parameter estimation and test reliability when the dependence was ignored. Results showed that the parameters of the new models were recovered well with the freeware WinBUGS. Fitting the new models to data without the person dependence did little harm to parameter estimation, but ignoring the person dependence by fitting standard IRT models yielded a shrunken scale and underestimated test reliability. We provided an empirical example of marital satisfaction to demonstrate the implications and applications of the new models.

Keywords Paired samples • Local person dependence • Multidimensional item response theory • Rasch models • Bayesian inference

Item response theory (IRT) models have been developed for categorical responses and have been widely applied in educational and psychological tests. It is common to calibrate item and person parameters with maximum likelihood estimation methods, in which the likelihood is simply the product of the likelihoods of all responses across items and persons. This method implies that all residuals are independent, given the model parameters. In practice, item residuals may be correlated across items (referred to as local item dependence; LID) when standard IRT models are fit. For example, LID may occur when tests consist of testlets or

K.-Y. Jin (✉)
Assessment Research Centre, The Hong Kong Institute of Education, Tai Po, Hong Kong
e-mail: kyjin@ied.edu.hk

W.-C. Wang
The Hong Kong Institute of Education, Tai Po, Hong Kong

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer
Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_9

item bundles (Wainer, Bradlow & Wang 2007; Wang & Wilson 2005), positively and negatively worded items (Wang, Chen & Jin 2015), and nonignorable missing data (Glas & Pimentel 2008; Holman & Glas 2005). In general, these studies have shown that the estimation of item and person parameters and test reliability is biased when LID exists but is neglected, and the newly developed, complicated IRT models can account for LID under these conditions.

Similar to LID, person residuals may be correlated when standard IRT models are fit (referred to as local person dependence; LPD). An example in which LPD occurs is cheating on examinations. When a group of examinees cheats on examinations, and standard IRT models are fit, the person residuals are no longer independent because these models do not consider cheating. Another example of LPD is paired-sample surveys where paired persons respond to something that is shared between them. For example, in a series of national surveys of marital satisfaction, a national sample of 50,379 married couples and 50,575 unmarried couples were recruited to respond to inventories on the strengths of marriage (Deal & Olson 2010). In the 2010 Married and Cohabiting Couples survey (http://www.bgsu.edu/ncfmr.html), researchers sampled 1075 nationally representative couples. In the first national survey of midlife development, siblings ($N = 950$) and twin pairs ($N = 1914$) were surveyed on the role of behavioral, psychological, and social factors in accounting for age-related variations in health and well-being (Pudrovska & Carr 2009). Unlike independent samples, paired samples are correlated and should not be treated as independent samples (Griffin & Gonzalez 1995; Kenny, Kashy & Cook 2006; Kenny, Mannetti, Pierro, Livi & Kashy 2002).

Paired samples cannot be analyzed with standard IRT models, in which all persons are assumed to be uncorrelated with one another. When LPD occurs but is ignored, the likelihoods are incorrect; therefore, the resulting parameter estimates will be biased, as when LID is ignored. To resolve this problem, new IRT models that consider LPD in paired samples are needed, which is the purpose of this study.

The remainder of the paper is organized as follows. First, standard IRT models are described. Second, new IRT models for LPD in paired samples are introduced. Third two simulation studies are conducted to evaluate the parameter recovery of the new models and the consequences of ignoring LPD, and their results are summarized. Fourth, the applications of the new models are demonstrated with an empirical example of marriage satisfaction. Finally, the conclusion and suggestions for future studies are provided.

1 Common IRT Models

Several IRT models have been developed for dichotomous or polytomous items. In the family of Rasch models, for example, the partial credit model (PCM; Masters 1982) can be expressed as follows:

$$\log\left(P_{nij}/P_{ni(j-1)}\right) = \theta_n - \delta_{ij}, \tag{1}$$

where P_{nij} and $P_{ni(j-1)}$ are the probabilities of scoring j and $j-1$, respectively, on item i for person n; θ_n is the latent trait for person n and is usually assumed to be normally distributed; and δ_{ij} is the jth difficulty of item i. When items are scored according to the same rubric (e.g., rating scale or Likert items), Eq. (1) can be constrained to share the same set of threshold parameters as follows:

$$\log\left(P_{nij}/P_{ni(j-1)}\right) = \theta_n - \left(\delta_i + \tau_j\right), \tag{2}$$

where δ_i is the overall difficulty of item i; τ_j is the jth threshold parameter for all items; and the other variables have already been defined. Eq. (2) is called the rating scale model (RSM; Andrich 1978).

When items are connected with the same stimulus (e.g., reading passages or figures), referred to as a testlet or an item bundle, and standard IRT models such as Eqs. (1) and (2) are fit, it is likely that item residuals within a testlet are correlated. If so, these IRT models are inappropriate, and the resulting parameter estimates are biased. To tackle the LID among items within a testlet, one can extend Eq. (1) (or (2)) by adding a testlet-specific random-effect parameter (Wang et al. 2015) as follows:

$$\log\left(P_{nij}/P_{ni(j-1)}\right) = \theta_n - \delta_{ij} + \gamma_{nd(i)}, \tag{3}$$

where $\gamma_{nd(i)}$is the random-effect parameter for person n on testlet d where item i is located, and the other parameters have already been defined. In this case, θ_n and $\gamma_{nd(i)}$are assumed to be normally and independently distributed. It is expected that adding the testlet-specific random-effect parameters would account for the dependence; thus, the model becomes correct, and the resulting parameter estimates are no longer biased.

In large-scale surveys, such as the Trends in International Mathematics and Science Study and Program for International Student Assessment, multistage sampling is often adopted. For example, a set of schools is sampled first, and students are then sampled from the chosen schools. Because students within the same school are often more homogenous in the measured latent trait than students from different schools, multilevel IRT models are needed to account for the clustering effect (Fox 2010). For example, Eq. (1) can be extended as follows:

$$\begin{aligned}
&\log\left(P_{nsij}/P_{nsi(j-1)}\right) = \theta_{ns} - \delta_{ij},\\
&\theta_{ns} = \theta_s + \varepsilon_{ns},\\
&\theta_s \sim N\left(0, \sigma_\theta^2\right),\\
&\varepsilon_{ns} \sim N\left(0, \sigma_\varepsilon^2\right),
\end{aligned} \tag{4}$$

where P_{nsij} and $P_{nsi(j-1)}$ are the probabilities of scoring j and $j-1$, respectively, on item i for student n in schools s, θ_{ns} is the latent trait of student n in school s, θ_s is the mean of the latent trait θ for students in school s, ε_{ns} is the residual, and the

other parameters have already been defined. When tests consist of testlet items, Eq. (4) can be further extended as follows:

$$\begin{aligned}
&\log\left(P_{nsij}/P_{nsi(j-1)}\right) = \theta_{ns} - \delta_{ij} + \gamma_{nsd(i)},\\
&\theta_{ns} = \theta_s + \varepsilon_{ns},\\
&\gamma_{nsd(i)} = \gamma_{sd(i)} + \varsigma_{nsd(i)},\\
&\theta_s \sim N\left(0, \sigma_\theta^2\right),\\
&\varepsilon_{ns} \sim N\left(0, \sigma_\varepsilon^2\right),\\
&\gamma_{sd(i)} \sim N\left(0, \sigma_{\gamma_d}^2\right),\\
&\varsigma_{nsd(i)} \sim N\left(0, \sigma_{\varsigma_d}^2\right),
\end{aligned} \tag{5}$$

where $\gamma_{nsd(i)}$ is the random-effect parameter for student n in school s on testlet d where item i is located; $\gamma_{sd(i)}$ is the mean of the random-effect parameters for students in school s on testlet d where item i is located; $\varsigma_{nsd(i)}$ is the residual; and the other parameters have already been defined. Jiao, Kamata, Wang, and Jin (2012) developed a submodel of Eq. (5) for dichotomous items as

$$\begin{aligned}
&\log\left(P_{ni1}/P_{ni0}\right) = \theta_{ns} - \delta_i + \gamma_{nd(i)},\\
&\theta_{ns} = \theta_s + \varepsilon_{ns},\\
&\theta_s \sim N\left(0, \sigma_\theta^2\right),\\
&\varepsilon_{ns} \sim N\left(0, \sigma_\varepsilon^2\right),\\
&\gamma_{nd(i)} \sim N\left(0, \sigma_{\gamma_d}^2\right),
\end{aligned} \tag{6}$$

where P_{nsi1} and P_{nsi0} are the probabilities of scoring 1 and 0, respectively, on item i for student n in schools s; and the other parameters have already been defined. Equation (6) is limited to dichotomous items, and the clustering effect is considered only at θ, leaving γ unconsidered. Although Eq. (5) considers the clustering effect among persons within the same cluster (school) and the testlet effect among items within the same testlet, this equation is not appropriate for paired samples, such as spouses or couples.

To the best of our knowledge, few IRT models exist for describing LPD. Cristante and Robusto (1999) and Robusto and Cristante (2010) developed the response dependence of subjects model (RDSM) to account for the dependence among persons in a small group. Unlike common IRT models in which the analysis unit is the individual person, the analysis unit in the RDSM is the individual cluster. Built upon a binomial model, the RDSM can be expressed as follows:

$$\log\left(P_{six}/P_{si(x-1)}\right) = \beta_s - \left[\eta_i + \lambda_s\left(2x - N_s - 1\right)\right], \tag{7}$$

where P_{six} and $P_{si(x-1)}$ are the probabilities of scoring x and $x-1$ on item i for cluster s, respectively, N_s is the size of cluster s, β_s is the location on the latent trait of cluster s, η_i is the location of item i, and λ_s is the dependence parameter of cluster s. The RDSM treats a total of N_s persons within cluster s as replications

of the same event, and the model assumes that persons in the same cluster have an equal probability of success on an item. The assumption of equal probability can be empirically tested. The value of λ_s determines the shape of the binomial distribution. When $\lambda_s = 0$, the distribution is uniform; when $\lambda_s > 0$, the distribution is unimodal; and when $\lambda_s < 0$, the distribution is concave. Therefore, the occurrence of person dependence within cluster s can be assessed with λ_s and the equal probability assumption. When λ_s is less than or equal to a limiting value (Cristante & Robusto 1999, p. 262), persons within cluster s are dependent; when λ_s is larger than the limiting value, one has to consider whether the equal probability assumption holds to determine person dependence.

Although it seems that the RDSM can account for LPD in small groups, the model has three major limitations. First, this model is limited to dichotomous items, but most psychological inventories use polytomous items. Second, although the equal probability assumption can be tested, the RDSM does not provide person measures for individual respondents, which has been criticized because individuals' scores in a cluster may not be simply aggregated to form a score for that cluster (Ganong 2003). Third, the RDSM assumes LPD is constant across items and thus fails to accommodate the variability of LPD across items. In this study, we developed a set of new IRT models to resolve these problems.

2 Development of New Models for Paired Samples

There are two kinds of paired samples. One is the natural pair, such as husband/wife, boyfriend/girlfriend, or older sibling/younger sibling, in which the relationship between the paired persons is natural. The other type of paired sample is the artificial pair in which the persons are paired to serve a purpose such as in experimental/control groups. In the examples of husband/wife, boyfriend/girlfriend, older sibling/younger sibling, and experimental/control groups, there is no doubt in assigning two persons in a pair to a reference group and a focal group. For example, husbands are assigned to one group, and wives to another. The membership assignment may not be possible in all situations, however, such as when identical twins or students randomly selected from schools are included. In this study, we focus on clear assignment.

Consider the husband/wife pair as an example. We can assign husbands to the reference group and wives to the focal group. The two groups could follow two different but correlated distributions. If so, standard IRT models such as the PCM (refer to Eq. (1)) in which all persons are assumed to follow the same distribution become inappropriate. To tackle this problem, we can extend the PCM as follows:

$$\text{Reference unit}: \quad \log\left(P_{sij}/P_{si(j-1)}\right)_{\mathrm{R}} = \theta_{s\mathrm{R}} - \delta_{ij}, \tag{8}$$

$$\text{Focal unit}: \quad \log\left(P_{sij}/P_{si(j-1)}\right)_{\mathrm{F}} = \theta_{s\mathrm{F}} - \delta_{ij}, \tag{9}$$

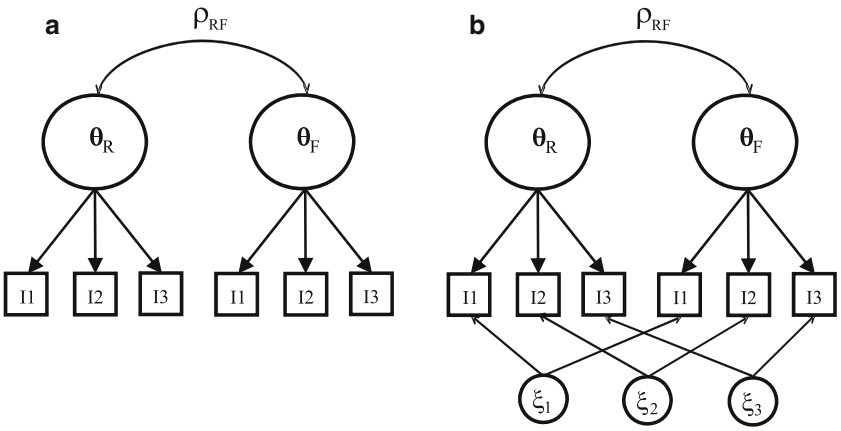


Fig. 1 Illustration of (**a**) the PCM-P, and (**b**) the PCM-PD in three items. *Note*: θ_R and θ_F are the latent traits of the focal and reference units; ρ_{RF} is the correlation between θ_F and θ_R; and ξ_1 to ξ_3 are the person dependence parameters on Items 1–3, respectively

where $\theta_{s\mathrm{F}}$ and $\theta_{s\mathrm{R}}$ are the latent traits for the focal and reference units in pair s, respectively; and δ_{ij} is the jth threshold parameter for item i. Equations (8) and (9) together are called the partial credit model for paired samples (PCM-P). When $\theta_{s\mathrm{F}}$ and $\theta_{s\mathrm{R}}$ follow an identical and independent distribution, the PCM-P reduces to the PCM. In other words, the PCM-P is an extension of the PCM in which the structure of the person pairs is specified (Fig. 1).

Following common practices in IRT, we can treat the item parameters in the PCM-P as fixed effects, and we can treat the person parameters $\boldsymbol{\theta}_s = [\theta_{s\mathrm{F}}, \theta_{s\mathrm{R}}]^{\mathrm{T}}$ as random effects following a bivariate normal distribution with mean $\boldsymbol{\mu} = [\mu_{\mathrm{F}}, \mu_{\mathrm{R}}]^{\mathrm{T}}$ and variance-covariance matrix $\Sigma = \begin{bmatrix} \sigma_{\mathrm{F}}^2 & \sigma_{\mathrm{FR}} \\ \sigma_{\mathrm{FR}} & \sigma_{\mathrm{R}}^2 \end{bmatrix}$. The correlation between the two latent traits, $\rho_{\mathrm{FR}} = \sigma_{\mathrm{FR}}/\sigma_{\mathrm{F}}\sigma_{\mathrm{R}}$, depicts the correlation in the latent traits between paired persons. If we treat the reference units as the pretest and the focal units as the posttest, then the PCM-P is equivalent to Andersen's (1985) longitudinal IRT model for repeated testings.

Although the PCM-P can describe the correlation in the latent trait between paired persons, LPD can still occur between paired persons, especially when the investigated latent traits involve an interpair relationship such as marital satisfaction and coping strategies in marriage. In responding to such inventories, perspective-taking may be engaged, which, in turn, may cause LPD. To account for LPD between paired persons, we extend Eqs. (8) and (9) as follows:

$$\text{Reference unit}: \quad \log\left(P_{sij}/P_{si(j-1)}\right)_{\mathrm{R}} = \theta_{s\mathrm{R}} - \delta_{ij} + \xi_{is}, \tag{10}$$

$$\text{Focal unit}: \quad \log\left(P_{sij}/P_{si(j-1)}\right)_{\mathrm{F}} = \theta_{s\mathrm{F}} - \delta_{ij} + \xi_{is}, \tag{11}$$

where ξ_{is} describes the item-level LPD on item i for pair s and is assumed to follow a normal distribution with mean 0 and variance $\sigma^2_{\xi i}$, and the other parameters have already been defined. A positive ξ_{is} increases the probability of endorsement, whereas a negative ξ_{is} decreases the probability. Equations (10) and (11) are called the partial credit model for dependence in paired samples (PCM-PD), which includes $2 + L$ random-effect parameters (L is the number of items). For model identification and ease of parameter interpretation, these random-effect parameters are assumed to be mutually independent. Because each ξ parameter is measured by a single item, precise estimates for individual persons are not possible. However, a precise estimate of $\sigma^2_{\xi i}$ is attainable with a sufficiently large sample. The magnitude of $\sigma^2_{\xi i}$ depicts the magnitude of LPD between paired persons for item i: the larger the variance, the larger the LPD. When $\sigma^2_{\xi i}$ is 0 for all items, the PCM-PD reduces to the PCM-P.

Adding ξ_{is} to δ_{ij}, the item parameters become random effects. Thus, the PCM-PD is actually a crossed random-effect IRT model (De Boeck 2008) because the persons and the items are random effects. Random item parameters have two implications. One is "random across items," which means the item parameters follow a random distribution. The other is "random across persons within items," which means the randomness in the item parameters is caused by the interaction between the persons and the items. The PCM-PD is more in line with the latter implication.

If LPD exists but is ignored by fitting a standard IRT model, the resulting scale will shrink, which also occurs when LID is ignored in testlets (Wang & Wilson 2005). Consider the illustration of a 4-point item with three thresholds of −2, 0, and 2. Conditional on θ, say, equal to 0, the probabilities for the four response categories are 0.06, 0.44, 0.44, and 0.06, respectively. Suppose there is a moderate LPD, say, $\sigma^2_{\xi_i} = 1$; then the marginal probabilities for the four response categories across the random effect will be 0.11, 0.39, 0.39, and 0.11, respectively. If we ignore the LPD and the threshold estimates, the marginal probabilities will be −1.27, 0, and 1.27, respectively. The scale has shrunk. As a consequence, σ^2_{F} and σ^2_{R} will be underestimated, and the difference between μ_{F} and μ_{R} will be attenuated.

3 Further Extensions

Sometimes, paired samples are randomly selected from different geographic units. For example, in the Married and Cohabiting Couples survey, a nationally representative sample of U.S. married and cohabiting adults was randomly selected based on the country geographic unit. Couples who come from the same country are likely to be more homogeneous than those from different countries. To account for this clustering effect for person pairs, Eqs. (10) and (11) can be extended as follows:

$$\text{Reference unit}: \ \log\left(P_{gsij}/P_{gsi(j-1)}\right)_{\mathrm{R}} = \theta_{gs\mathrm{R}} - \delta_{ij} + \xi_{gis}, \ \theta_{gs\mathrm{R}} = \theta_{g\mathrm{R}} + \varepsilon_{gs\mathrm{R}}, \tag{12}$$

$$\text{Focal unit}: \log\left(P_{gsij}/P_{gsi(j-1)}\right)_{\text{F}} = \theta_{gs\text{F}} - \delta_{ij} + \xi_{gis},\ \theta_{gs\text{F}} = \theta_{g\text{F}} + \varepsilon_{gs\text{F}}, \quad (13)$$

where g refers to level-2 group membership (e.g., country), and s refers to level-1 pair membership; $\theta_{g\text{R}}$ and $\theta_{g\text{F}}$ are assumed to follow a bivariate normal distribution; ξ_{gis} describes the item-level LPD on item i for pair s in group g and is assumed to follow a normal distribution with mean 0 and variance $\sigma^2_{\xi i}$; and $\varepsilon_{gs\text{R}}$ and $\varepsilon_{gs\text{F}}$ are assumed to follow another bivariate normal distribution. Furthermore, when necessary, covariates can be incorporated into this multilevel model to account for θ and ξ to form explanatory IRT (De Boeck & Wilson 2004).

In these models, the target latent trait θ is unidimensional. Where appropriate, it can be generalized to be multidimensional. For example, a test often consists of multiple subtests, and each subtest measures a distinct latent trait. In such cases, all subtests can be analyzed jointly with a multidimensional approach. It has been shown that multidimensional approaches are more efficient than consecutive unidimensional approaches where each subtest is analyzed separately, one subtest at a time (Adams, Wilson, & Wang 1997; Briggs & Wilson 2003). To analyze all subtests jointly, we can extend Eqs. (10) and (11) as

$$\text{Reference unit}: \log\left(P_{dsij}/P_{dsi(j-1)}\right)_{\text{R}} = \theta_{ds\text{R}} - \delta_{dij} + \xi_{dis}, \quad (14)$$

$$\text{Focal unit}: \log\left(P_{dsij}/P_{dsi(j-1)}\right)_{\text{F}} = \theta_{ds\text{F}} - \delta_{dij} + \xi_{dis}, \quad (15)$$

where subscript d ($d = 1, \ldots, D$) is the index of subtests, and the other parameters have already been defined. Now, $\boldsymbol{\theta} = [\theta_{1\text{F}}, \ldots, \theta_{D\text{F}}, \theta_{1\text{R}}, \ldots, \theta_{D\text{R}}]^{\text{T}}$ contains $2 \times D$ elements and is assumed to follow a multivariate normal distribution.

Although our models have no item slope parameters, these parameters can be easily added where appropriate. In this study, we focus on the PCM-P and the PCM-PD (see Eqs. (8)–(11)), and leave the multilevel model (see Eqs. (12) and (13)) and the multidimensional model (see Eqs. (14) and (15)) for future studies. Although designed specifically for pairs, the new models can be easily generalized to more than two persons in a cluster as long as the membership within a cluster can be specified. For example, in a family survey, all family members are surveyed, including the father, mother, first-born child, second-born child, and so on; in an experiment with matched samples, there may be multiple experimental groups or multiple control groups. In contrast, sometimes the membership within a pair or cluster cannot be specified, such as with identical twins or students within a school; therefore, the PCM-P and the PCM-PD are not applicable. In such cases, Eq (4) or standard multilevel IRT models can be used.

4 Parameter Estimation

The PCM-PD has as many as $2 + L$ dimensions, which makes marginal maximum likelihood estimation methods infeasible. We thus adopted the Bayesian approach with Markov chain Monte Carlo (MCMC) methods for parameter estimation. In the Bayesian approach, a statistical model and the prior distributions of the model parameters are specified to yield a joint posterior distribution. After a sequential sampling, the posterior distribution of each parameter is formed. Its mean and standard deviation can be reported as the point estimate and corresponding standard error of the parameters, respectively. The freeware WinBUGS (Spiegelhalter, Best, Carlin & Linde 2002) was used in this study.

In the following simulation studies and empirical examples, the target latent traits $\boldsymbol{\theta}_s = [\theta_{s\mathrm{R}}, \theta_{s\mathrm{F}}]^{\mathrm{T}}$ were assumed to follow a bivariate normal distribution. The mean of $\theta_{s\mathrm{R}}$ (i.e., μ_{R}) was set at 0 for model identification, so the mean of $\theta_{s\mathrm{F}}$ (i.e., μ_{F}), the variance-covariance matrix of $\boldsymbol{\theta}_s$ (i.e., $\boldsymbol{\Sigma}$), the item thresholds (i.e., δ-parameters), and the item-level person dependence (i.e., ξ-parameters) were freely estimated. The following priors were given in WinBUGS: $N(0, 10)$ for δ and μ_{F}, Gamma $(0.1, 0.1)$ for the inverse of $\sigma^2_{\xi\mathrm{i}}$, and a Wishart $(\boldsymbol{\Psi}, 2)$ for the inverse of $\boldsymbol{\Sigma}$, where $\boldsymbol{\Psi}$ is a 2×2 identity matrix.

5 Simulation Studies

5.1 Design

Two simulation studies were conducted. Study 1 aimed to examine parameter recovery by fitting the PCM-PD to the PCM-PD data and to evaluate the consequences of ignoring the dependence between paired persons by fitting the PCM-P to the PCM-PD data. Item responses were generated according to the PCM-PD. The sample size was either 500 or 1000 pairs, and the test consisted of 10 four-point rating scale items. The latent traits for the paired samples were generated from a bivariate normal distribution with mean $\boldsymbol{\mu} = \left[\mu_{\mathrm{R}}, \mu_{\mathrm{F}}\right]^{\mathrm{T}} = \left[0, 0.5\right]^{\mathrm{T}}$ and variance-covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}$. The item thresholds were set between -2 and 2. The values of $\sigma^2_{\xi\mathrm{i}}$ were set between 0.4 and 1.2, ranging from small to large. One hundred replications were completed in each condition. It was expected that the parameter estimation would be fairly good when the data-generating PCM-PD was fit to the PCM-PD data but would be biased when the PCM-P was fit to the PCM-PD data.

Study 2 aimed to investigate the parameter recovery of the PCM-P and the consequences of fitting an unnecessarily complicated PCM-PD to data simulated from the PCM-P. The data were simulated from the PCM-P and analyzed with the PCM-P and the PCM-PD. The settings were similar to those in Study 1. A total

of 100 replications were conducted. It was expected that the PCM-P would have a good parameter recovery, and the PCM-PD would yield estimates for $\sigma^2_{\xi i}$ very close to 0 and estimates for the other parameters very close to their true values. In other words, it would do little harm to fit the PCM-PD to the PCM-P data.

5.2 Analysis

The first 5000 iterations were discarded for burn-in, and the second 5000 iterations were retained. Afterward, parameter estimates were sampled from the remaining iterations per every 10 values. The bias and root mean square error (RMSE) of the parameter estimates were computed to evaluate the parameter recovery. In addition, the squared correlation between the estimates and the true values of the person measures was computed to illustrate the consequence of ignoring the LPD on test reliability. The Akaike information criterion (AIC), Bayesian information criterion (BIC), and deviance information criterion (DIC) were used to compare the models. The DIC is a Bayesian version of the AIC and BIC. Like the AIC and BIC, the DIC considers both the measure of model adequacy and the measure of model complexity. The DIC is based on the posterior distribution of the log-likelihood and can be obtained from WinBUGS.

5.3 Results

Study 1. The upper panel of Table 1 summarizes the bias and RMSE values when the PCM-P and the PCM-PD were fit to the PCM-PD data. As expected, the PCM-P yielded poor estimation, whereas the PCM-PD recovered the parameters very well. When the PCM-P was fit, and there were 500 couples, the bias was between −0.799 and 0.704, and the RMSE was between 0.089 and 0.807; when there were 1000 couples, the bias was between −0.916 and 1.020, and the RMSE was between 0.100 and 1.025. It seemed the parameter recovery for the PCM-P could not be improved by adopting a large number of couples. As expected, the PCM-P underestimated the θ variances, suggesting the scales were noticeably shrunken compared to their true values. Because μ_F was underestimated, the difference between μ_F and μ_R was underestimated. Note that the correlation between the reference and focal units (ρ_{FR}) was precisely estimated in the PCM-P.

In contrast, all the estimated parameters were recovered accurately in the PCM-PD. When there were 500 couples, the bias was between −0.032 and 0.064, and the RMSE was between 0.061 and 0.245; when there were 1000 couples, the bias was between −0.026 and 0.041, and the RMSE was between 0.044 and 0.204. It seemed that the larger the number of couples (sample size), the better the parameter estimation.

Table 1 Parameter recovery summary for the PCM-P and the PCM-PD in simulation studies 1 and 2

	PCM-P				PCM-PD			
	$N = 500$		$N = 1000$		$N = 500$		$N = 1000$	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
Study 1								
Threshold								
Max	0.704	0.807	1.020	1.025	0.018	0.207	0.013	0.158
Min	−0.799	0.089	−0.916	0.100	−0.032	0.100	−0.026	0.060
Mean	−0.109	0.419	0.051	0.411	−0.004	0.133	−0.005	0.095
σ^2_ξ								
Max	–	–	–	–	0.064	0.245	0.041	0.204
Min	–	–	–	–	−0.002	0.127	−0.004	0.072
Mean	–	–	–	–	0.031	0.178	0.013	0.123
Distribution								
μ_F	−0.171	0.176	−0.172	0.174	−0.002	0.062	−0.008	0.044
σ^2_F	−0.548	0.550	−0.543	0.544	0.030	0.113	0.006	0.070
σ^2_R	−0.560	0.561	−0.549	0.550	−0.008	0.088	0.007	0.074
σ_{FR}	−0.131	0.134	−0.126	0.128	0.002	0.061	0.005	0.046
ρ_{FR}	0.079	0.094	0.083	0.091	−0.002	0.051	0.003	0.037
Study 2								
Threshold								
Max	0.021	0.176	0.007	0.112	0.182	0.261	0.137	0.168
Min	−0.045	0.083	−0.022	0.058	−0.188	0.087	−0.119	0.059
Mean	−0.003	0.112	−0.003	0.077	−0.006	0.142	0.023	0.102
σ^2_ξ								
Max	–	–	–	–	0.116	0.124	0.084	0.087
Min	–	–	–	–	0.091	0.094	0.071	0.073
Mean	–	–	–	–	0.100	0.105	0.077	0.080
Distribution								
μ_F	0.007	0.052	−0.002	0.043	0.042	0.070	0.029	0.054
σ^2_F	0.015	0.093	0.008	0.062	0.151	0.183	0.118	0.135
σ^2_R	0.021	0.094	0.004	0.062	0.163	0.194	0.113	0.132
σ_{FR}	0.000	0.056	0.004	0.039	0.031	0.071	0.030	0.052
ρ_{FR}	−0.004	0.051	0.003	0.033	−0.014	0.053	−0.005	0.033

Note: N denotes the number of couples;–denotes not applicable

In terms of test reliability, the mean estimates in the PCM-P were .784 and .776 when there were 500 and 1000 couples, respectively; the mean estimates in the PCM-PD were .794 and .787 when there were 500 and 1000 couples, respectively. It appeared that ignoring the dependence tended to underestimate the test reliability slightly.

For the 100 replications, the AIC, BIC, and DIC always favored the data-generating PCM-PD, suggesting that these fit indices were very powerful in

selecting the true model. A high power was meaningful only when the Type I error rate could be well controlled. In simulation study 2, we investigated whether these indices could yield well-controlled Type I error rates.

Study 2. The lower panel of Table 2 summarizes the bias and RMSE values when the PCM-P and the PCM-PD were fit to the PCM-P data. Both models yielded similar parameter estimates. In the PCM-PD, all the estimates of $\sigma^2_{\xi_i}$ were close to their expected value of 0. For example, the estimates of $\sigma^2_{\xi_i}$ were between 0.091 and 0.116 when there were 500 couples, and between 0.071 and 0.084 when there were 1000 couples. In the PCM-PD, σ^2_{F} and σ^2_{R} were slightly overestimated. Both models yielded the same test reliability. The mean estimates of test reliability in both models were between .858 and .859 when there were 500 and 1000 couples. Thus, fitting a complicated model (i.e., the PCM-PD) to data without person dependence did little harm.

Contradictory to our expectations, the AIC, BIC, and DIC always favored the PCM-PD. This inflated Type I error rate might be due to the large number of random-effect parameters in the PCM-PD. Thus, it would be more appropriate to evaluate the magnitudes of $\sigma^2_{\xi_i}$ when selecting a model. From the simulation results, it seemed that a variance of 0.1 could be treated as a cut point to indicate whether the person dependence existed.

6 An Empirical Example of Marriage Satisfaction

The data were drawn from the Familial Responses to Financial Instability project (National Center for Family & Marriage Research, Diamond & Hicks 2010) in which 630 couples were recruited in the United States. A scale measuring marriage satisfaction consisted of 14 five-point items: $0 = \textit{not at all}$, $1 = \textit{a little}$, $2 = \textit{moderately}$, $3 = \textit{quite a bit}$, and $4 = \textit{extremely}$. Six items were negatively worded and thus were reversely recoded. A higher score showed more satisfaction in the marriage. The PCM-P and the PCM-PD were fit to the data. Wives were treated as the reference units. The posterior predictive p-value of the Bayesian chi-square was computed to check the model-data fit. Additionally, the Q_3 statistic (Yen 1984) of the person residuals within couples was computed to assess the magnitude of LPD as follows:

$$h_{ni} = Y_{ni} - E_i\left(\widehat{\theta}_n\right), \tag{16}$$

$$Q_3 = cor\left(h_n, h_{n'}\right), \tag{17}$$

where Y_{ni} is the observed score of person n on item i; $E_i\left(\widehat{\theta}_n\right)$ is the expected score of person n on item i with an ability estimate of $\widehat{\theta}_n$; h_{ni} is the residual of person n on item i; and $cor\left(h_n, h_{n'}\right)$ is the correlation between the two sets of residual scores of

Table 2 Estimates for the Dependence Effect in Real Couples and Randomly Paired Couples

No.	Item	Real couples		Random couples	
		$\widehat{\sigma}^2_\xi$	SE	$\widehat{\sigma}^2_\xi$	SE
1	How well does your partner meet your needs?	0.139	0.076	0.053	0.027
2	How satisfied are you with your relationship?	0.064	0.032	0.040	0.016
3	How much do you wish you hadn't gotten into this relationship?	1.014	0.250	0.718	0.213
4	How much do you love your partner?	0.104	0.058	0.071	0.042
5	How many problems are there in your relationship?	1.960	0.374	0.368	0.134
6	I don't often worry about my partner abandoning or getting too close to me.	5.157	0.565	2.449	0.264
7	I am somewhat uncomfortable being close to my partner.	0.876	0.230	0.435	0.147
8	I find that my partner is reluctant to get as close as I would like.	1.077	0.232	1.029	0.210
9	To what extent have you felt close/connected in your interactions with your partner in the last 6 months?	0.154	0.085	0.043	0.021
10	How much do you feel that you and your partner share and are working together toward your financial goals for the future?	1.774	0.292	0.258	0.107
11	How much do you feel that you and your partner share and are working together toward other goals for the future?	0.799	0.162	0.086	0.046
12	How likely is it that you will still be in this relationship 1 year from now?	0.292	0.227	0.150	0.077
13	How much do you think about leaving this relationship?	0.125	0.069	0.059	0.027
14	How much do financial matters influence whether you stay in this relationship?	3.002	0.443	1.508	0.263

Note: Items 3, 5, 7, 8, 13, and 14 are negatively worded items and thus reversely recoded in the analyses

person n and person n' across items. Under the null hypothesis of no LPD, Q_3 will be approximately normally distributed with mean $-1/(N \times 2 - 1)$ (N is the number of pairs) and standard deviation $\sqrt{1/T - 2}$ (T is the number of items). Because there were 630 couples and 14 items, if all person residuals were independent, the mean and standard deviation of Q_3 would be -0.001 and 0.289, respectively.

The posterior predictive p-value for the PCM-P and the PCM-PD was .004 and .470, respectively, indicating that only the PCM-PD fit the data well. The empirical mean Q_3 was 0.007 for the PCM-P and the PCM-PD; the empirical standard deviation was 0.364 for the PCM-P and 0.333 for the PCM-PD. It seemed that the PCM-PD had a better fit. Table 2 summarizes the estimates for σ^2_ξ of the 14 items in the PCM-PD. It appeared that the person dependence was strong for a few items. The two items with the largest variances were Item 6 ("I don't often worry about my partner abandoning or getting too close to me."; $\sigma^2_\xi = 5.16$) and Item 14 ("How much do financial matters influence whether you stay in this relationship?"; $\sigma^2_\xi = 3.00$). The two items with the smallest variances were Item 2 ("How satisfied are you with your relationship?"; $\sigma^2_\xi = 0.06$) and Item 13 ("How much do you love your partner?"; $\sigma^2_\xi = 0.10$). The large person dependence in Items 6 and 14 might be because these two items required more engagement in perspective-taking between husbands and wives (Galinsky & Moskowitz 2000).

The estimates of μ_F, σ^2_F, σ^2_R, and σ_{FR} in the PCM-P were 0.16, 2.27, 2.23, and 1.53, respectively, and 0.28, 5.37, 5.29, and 3.61 in the PCM-PD, respectively. The estimate of ρ_{FR} was .68 for both the PCM-P and the PCM-PD, which suggested a moderate correlation in marriage satisfaction between couples. As demonstrated in the simulations, the scale in the PCM-P shrunk substantially. The Pearson correlation coefficient between sampled person measures from the MCMC draws was computed as an estimate of test reliability. The test reliability for husbands and wives in the PCM-P was .867 and .874, respectively, and in the PCM-PD, it was .882 and .889, respectively. Consistent with the simulations, the test reliability was slightly underestimated by the PCM-P.

Finally, we adopted the random-pairing strategy (Thiessen, Young, & Delgado 1997) to illustrate the correlation between husbands and wives and the interaction effect. A replicated dataset was created in which husbands and wives were randomly paired. The PCM-PD was then fit to the new dataset. The results showed that the estimate of ρ_{FR} was as small as .035, and the estimate of σ^2_ξ (listed in the right panel of Table 2) was between 0.040 and 2.449, which were much smaller than those for the real couples (listed in the left panel of Table 2). Thus, the large correlation and dependence found between husbands and wives in the real couples were not due to random errors.

7 Conclusion and Discussion

When paired samples are surveyed, their responses are likely dependent on each other. When standard IRT models are fit to such data, the person residuals within a pair may be correlated. If so, the parameter estimates will be biased, making the subsequent decisions erroneous. To resolve this problem, we developed the PCM-P and the PCM-PD to account for the correlation in the latent trait and the dependence between paired persons.

Two simulation studies were conducted to assess the parameter recovery of the PCM-P and the PCM-PD when the data-generating models were fit to the simulated data and to assess the consequences for the parameter estimation of misfitting the PCM-P to the PCM-PD data and the PCM-PD to the PCM-P data. Results demonstrated that the parameters of the PCM-P and the PCM-PD could be recovered fairly well with WinBUGS. Fitting the unnecessarily complicated PCM-PD to the PCM-P data did little harm to parameter estimation and yielded close to 0 estimates for σ_{ξ}^{2}. In contrast, ignoring the dependence by fitting the PCM-P to the PCM-PD data led to a shrunken scale, biased estimates for the item parameters, and underestimated test reliability. An empirical example of couples' marriage satisfaction was provided to demonstrate the implication and applications of the PCM-P and the PCM-PD. The correlation in marital satisfaction between husbands and wives was positive and moderate, and the dependence was large in a few items.

Several issues require further investigation. First, the slight overestimation in test reliability when LPD is ignored contradicts the underestimation in testlet response models (Wainer & Lukhele 1997; Wainer & Wang 2000). The two possible reasons for the inconsistency are the use of different models to account for different kinds of dependence (LPD vs. LID) and the presence of scale shrinkage. In the computation of (marginal) test reliability, test information is integrated over person distribution. If the scale remains unchanged when the dependence is ignored, then the marginal test reliability will be overestimated. However, because the scale shrinks to some degree when the dependence is ignored, the marginal test reliability will be affected by the scale shrinkage to an unknown degree, so it can be underestimated or overestimated. This issue should be further investigated.

The second issue that requires further investigation is the performance and applications of Eqs. (12)–(15), which need to be assessed and demonstrated. Third, the person dependence in this study was treated as a nonrecursive (nondirectional) process, which implies that a person can influence the other person in the same pair, and vice versa. In some cases, the process can be recursive (e.g., a parent can affect his or her child, but not vice versa), which calls for a new set of IRT models for the recursive effect.

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23.

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50*, 3–16.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–573.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*, 87–100.

Cristante, F., & Robusto, E. (1999). Assessing dependence among subjects' responses. *Mathematical Social Sciences, 38*, 259–274.

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*, 533–559.

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

Deal, R. L., & Olson, D. H. (2010). *The remarriage checkup: Tools to help your marriage last a lifetime*. Ada, MI: Bethany House.

Fox, J. P. (2010). *Bayesian item response modeling*. New York: Springer.

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology, 78*, 708–724.

Ganong, L. H. (2003). Selecting family measurements. *Journal of Family Nursing, 9*, 184–206.

Glas, C. A. W., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement, 68*, 907–922.

Griffin, D., & Gonzalez, R. (1995). Correlational analysis of dyad-level data in the exchangeable case. *Psychological Bulletin, 118*, 430–439.

Holman, R., & Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology, 58*, 1–17.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement, 49*, 82–100.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York, NY: Guilford Press.

Kenny, D. A., Mannetti, L., Pierro, A., Livi, S., & Kashy, D. A. (2002). The statistical analysis of data from small groups. *Journal of Personality and Social Psychology, 83*, 126–137.

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

National Center for Family and Marriage Research, Diamond, L., & Hicks, A. (2010). *Familial responses to financial instability, "It's all your fault": Predictors and implications of blame in couples under economic strain, 2009 [United States]*. ICPSR26544-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2010-05-20.

Pudrovska, T., & Carr, D. (2009). Age at first birth and fathers' subsequent health: Evidence from sibling and twin models. *American Journal of Men's Health, 3*, 104–115.

Robusto, E., & Cristante, F. (2010). The parameterization and the analysis of small groups by means of the response dependence of subjects model. *European Psychologist, 15*, 91–98.

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, 64*, 583–639.

Thiessen, D., Young, R. K., & Delgado, M. (1997). Social pressures for assortative mating. *Personality and Individual Differences, 22*, 157–164.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.

Wainer, H., & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational and Psychological Measurement, 57*, 749–766.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*, 203–220.

Wang, W.-C., Chen, H.-F., & Jin, K.-Y. (2015). Item response theory models for wording effects in mixed-format scales. *Educational and Psychological Measurement, 75*, 157–178.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*, 126–149.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125–145.

Using Sample Weights in Item Response Data Analysis Under Complex Sample Designs

Xiaying Zheng and Ji Seung Yang

Abstract Large-scale assessments are often conducted using complex sampling designs that include the stratification of a target population and multi-stage cluster sampling. To address the nested structure of item response data under complex sample designs, a number of previous studies proposed the multilevel/multidimensional item response models. However, incorporating sample weights into the item response models has been relatively less explored. The purpose of this study is to assess the performance of four approaches to analyzing item response data that are collected under complex sample designs: (1) single-level modeling without weights (ignoring complex sample designs), (2) the design-based (aggregate) method, (3) the model-based (disaggregate) method, and (4) the hybrid method that addresses both the multilevel structure and the sampling weights. A Monte Carlo simulation study is carried out to see whether the hybrid method can yield the least biased item/person parameter and level-2 variance estimates. Item response data are generated using the complex sample design that is adopted by PISA 2000, and bias in estimates and adequacy of standard errors are evaluated. The results highlight the importance of using sample weights in item analysis when a complex sample design is used.

Keywords Complex sample design • Multilevel item response theory • Sample weights • Pseudo maximum likelihood

1 Introduction

Large-scale educational assessments are often conducted through complex sampling designs for the purpose of reducing costs or improving precision for subgroup analyses relative to simple random sampling (SRS) (Heeringa, West & Berglund 2010). Such design typically includes a stratification of target population and multi-stage cluster sampling within each stratum that result in unequal selection

X. Zheng • J.S. Yang (✉)
College of Education, University of Maryland, 1230 Benjamin Building, College Park, MD 20742, USA
e-mail: xyzheng@umd.edu; jsyang@umd.edu

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_10

probabilities for different clusters and/or subjects within clusters. Data collected under complex sampling designs have a multilevel structure and sampling weights for units at each level. While traditional item response theory (IRT) models usually assume that examinees are independent and identically distributed across clusters (e.g. schools), these assumptions seldom hold for large-scale assessments that utilize complex sampling designs.

To address the nested structure of item response data under the complex sample designs, a number of previous studies proposed the model-based multilevel item response models (e.g., Adams, Wilson & Wu 1997; Fox & Glas 2001; Jiao, Kamata, Wang & Jin 2012; Kamata 2001), where clustering effects are treated as random effects. Multilevel IRT models have gained popularity in recent years as it addresses the person clustering that is common in education settings. But the sample weights are often not considered in estimating multilevel IRT models.

The second method to analyze complex sample data is design based, which incorporates the complex sample weights into likelihood, resulting in pseudolikelihood for point estimation (see e.g., Binder 1983; Skinner 1989). Taylor Series linearization, jackknifing or balanced repeated replication (BRR) methods are utilized for standard error estimation (see e.g., Rust 1985). However, the design-based method has been less explored in the context of item response models. One example of applying design-based method to IRT models is the work by Mislevy, Beaton, Kaplan, and Sheehan (1992), where a two-stage plausible value method is used to deal with sparse matrix of item responses. In stage 1, a unidimensional IRT calibration is conducted to obtain item parameters through marginal likelihood estimation. In stage 2, multiple imputations (Rubin 1987) of latent scores (also known as plausible values) are conducted via a latent regression model that treats item parameters from stage 1 as fixed. Sample weights are incorporated to the stage 2 model in a design-based manner to estimate parameters and standard errors. The plausible value method provides a practical framework for handling complex samples, and allows convenience for secondary data users. Another example of using design-based method in IRT modeling was explored by Cai (2013), which demonstrates that the sampling weights could be incorporated into one-level multiple-group IRT models to obtain more accurate population-level inferences.

The third approach to dealing with complex sample data combines the model-based and design-based methods by incorporating complex sampling weights in the likelihood of multilevel models. For standard errors, sandwich estimators can be used. The method has previously been evaluated in linear multilevel model (Pfeffermann, Skinner, Holmes, Goldstein & Rasbash 1998) and multilevel logistic regression (Rabe-Hesketh & Skrondal 2006), and has shown superior performance in reducing bias in point estimates. Rabe-Hesketh and Skrondal (2006) characterize this method as "a hybrid aggregated-disaggregated approach". We use "hybrid method" to refer to this combined approach throughout the manuscript. The hybrid method has also been examined using the data of the Trends in International Mathematics and Science Study (TIMSS) in the context of linear multilevel modeling (Laukaityte 2013). As far as the authors are aware of, the hybrid method has never been explored in IRT models.

The purpose of this paper is to assess the performance of four approaches to analyzing item response data that are collected under complex sample designs: (1) single-level IRT without weights, (2) the model-based method (multilevel IRT without weights), (3) the design-based method (single-level IRT with weights), and (4) the hybrid method (multilevel IRT with weights). We are particularly interested in seeing whether the hybrid method can yield the least biased item parameters and level-2 variance estimates under different conditions. To do so, we first briefly introduce complex sampling designs. A multilevel unidimensional model is then described. The marginal pseudolikelihood for the model is presented. The sandwich estimator for standard error estimation is also introduced. Finally a Monte Carlo simulation study is carried out to examine the performance of the pseudo-maximum-likelihood method in comparison with traditional design-based and model-based methods. Bias in estimates and adequacy of standard errors are evaluated across these methods.

Large-scale assessment data are routinely collected with complex sample designs. But the sample weights are often ignored in item analysis, which might lead to biased item parameter estimates and misleading inference on the target finite population. The results of the study highlight the importance of using sample weights in item analysis when a complex sample design is used.

2 Complex Sample Weights

In large-scale tests such as Programme for International Student Assessment (PISA), it is usually not practical to conduct simple random sampling (SRS) on the student level directly. Instead a complex sampling design is implemented to obtain student samples. This paper will keep using the terms "schools" and "students" for illustrative purpose.

Let's consider a complex case of cluster sampling, where stratification is carried out at both levels. The following indices are used:

- $h = 1, \ldots, H$ is the index for stratum at the school level.
- $k = 1, \ldots, K_h$ is the index for school within school-level stratum h.
- $g = 1, \ldots, G_{kh}$ is the index for within-school stratum of school k that is in school-level stratum h.
- $j = 1, \ldots, J_{gkh}$ is the index for student who is from within-school stratum g of school k, where school k is from school-level stratum h.

All schools are first separated to H school-level strata according to some grouping variables (e.g., public or private status and proportion of minority students). Let A_h and a_h be the total number of schools in stratum h and the number of schools to be sampled in stratum h, respectively. Suppose that schools in stratum 1 are oversampled compared to schools in stratum 2. Then a_1 and a_2 are decided in such a way that $a_1/A_1 > a_2/A_2$.

Within stratum h, a two-stage sampling is carried out, where schools are sampled in the first stage, and students are then selected from each sampled school on the second stage. A common way to conduct the first-stage sampling is through Probability Proportional to Size (PPS) sampling (see e.g., Kish 1965). With PPS, the probability of a school k being sampled is proportional to the total number of students in this school, N_{kh}. Let N_h be the population of students in stratum h. Then the selection probability for school k can be written as:

$$P_{k|h} = a_h \times N_{kh}/N_h. \tag{1}$$

The level-2 weights $W_{k|h}$ is the inverse of $P_{k|h}$.

In the second stage, the stratified random sampling is implemented. Students are further stratified within each school to G groups based on some student-level grouping variables (e.g., ethnicity). Students are then randomly selected from each group. Within school k in stratum h, let N_{gkh} and n_{gkh} be the total number of students in group g, and the number of students to be sampled in group g respectively. Suppose students in group 1 are over-sampled compared to students in group 2. Then n_{1kh} and n_{2kh} are decided in such a way so that $n_{1kh}/N_{1kh} > n_{2kh}/N_{2kh}$.

The conditional selection probability of student j in group g given that his/her school has already been selected is written as:

$$P_{j|g,k,h} = n_{gkh}/N_{gkh}. \tag{2}$$

The level-1 conditional weight $W_{j|gkh}$ is the inverse of $P_{j|gkh}$.

The overall unconditional probability of a student being selected is:

$$P_{jgkh} = P_{k|h} \times P_{j|g,k,h} = a_h \times N_{kh}/N_h \times n_{gkh}/N_{gkh}. \tag{3}$$

As a result, all the students in the same group g, school k, stratum h would have the same overall unconditional selection probability, while schools and students across different strata would have different weights.

3 Multilevel IRT Model and Pseudolikelihood

For illustration purpose, this section describes a two-level 2-parameter logistic IRT model. The marginal pseudolikelihood of the model as well as the sandwich estimator for standard errors are also presented. The IRT model and its estimation could easily be extended to polytomous or mixed item types, and situations with more than two levels.

3.1 Multilevel IRT Model

Let y_{ijk} be the observed response to item i, $(i = 1, \ldots, I)$ for student j in school k. Then θ_{jk}, the latent score for student j in school k, can be expressed as the sum of school level latent mean $\xi_{.k}$ and the individual deviation score δ_{jk}. In a dichotomous two-level unidimensional IRT model, let α_i be the slope on the latent variables at both level 1 and level 2 for cross-level measurement invariance assumption. β_i is the intercept for item i. The conditional likelihood of student j from school k answering item i correctly is:

$$f_{ijk} = f(y_{ijk} = 1 \mid \xi_{.k}, \delta_{jk}) = \frac{1}{1 + \exp(-\beta_i - \alpha_i \xi_{.k} - \alpha_i \delta_{jk})}. \tag{4}$$

3.2 Conventional Likelihood

If we do not consider the complex sample weights, the conditional density for an observed response y_{ijk} is:

$$f_{\boldsymbol{\lambda}}(y_{ijk} \mid \xi_{.k}, \delta_{jk}) = f_{ijk}^{y_{ijk}} (1 - f_{ijk})^{1 - y_{ijk}}, \tag{5}$$

where $\boldsymbol{\lambda}$ is a vector of parameters to be estimated. The contribution of a student's responses across all items to the marginal likelihood, conditional on level-2 random effect of school k is:

$$L_{j|k} = \int \prod_{i=1}^{I} f_{\boldsymbol{\lambda}}(y_{ijk} \mid \xi_{.k}, \delta_{jk}) g_1(\delta_{jk}) d\delta_{jk}, \tag{6}$$

where $g_1(\delta_{jk})$ is the distribution of level-1 latent variable δ_{jk}. The contribution of a level-2 school k to the marginal likelihood is:

$$L_k = \int \prod_{j=1}^{J_k} L_{j|k} g_2(\xi_{.k}) d\xi_{.k}, \tag{7}$$

where $g_2(\xi_{.k})$ is the distribution of level-2 latent variable $\xi_{.k}$. The marginal likelihood of the model to be maximized to obtain parameter estimates is the product of each school's contribution to the marginal likelihood:

$$L = \prod_{k=1}^{K} L_k. \tag{8}$$

3.3 Pseudolikelihood

Let $W_{k|h}$ be the conditional weight for school k in stratum h and $W_{j|g,k,h}$ be the conditional level-1 weight for student j in within-school stratum g, given that his/her school k has already been selected in the first stage. The contribution of student j to the marginal pseudolikelihood conditional on level-2 random effect can be obtained by rewriting Eq. (6) with weights as:

$$L^*_{j|gkh} = \int \left[\prod_{i=1}^{I} f_{\lambda}(y_{ijk} \mid \xi_{.k}, \delta_{jk})^{W_{j|g,k,h}} \right] g_1(\delta_{jk}) d\delta_{jk}. \tag{9}$$

And the contribution of school k in stratum h to the marginal pseudolikelihood can be written as:

$$L^*_{k|h} = \int \left[\prod_{g=1}^{G_{kh}} \prod_{j=1}^{J_k} \left(L^*_{j|gkh} \right)^{W_{k|h}} \right] g_2(\xi_{.k}) d\xi_{.k}. \tag{10}$$

Finally, the likelihood of the model is:

$$L^* = \prod_{h=1}^{H} \prod_{k=1}^{K_h} L^*_{k|h} = \prod_{h=1}^{H} \prod_{k=1}^{K_h} \int \left[\prod_{g=1}^{G_{kh}} \prod_{j=1}^{J_k} \left(L^*_{j|gkh} \right)^{W_{k|h}} \right] g_2(\xi_{.k}) d\xi_{.k}. \tag{11}$$

Thus, weights are incorporated into the likelihood of the multilevel model to replicate units at both levels. As Rabe-Hesketh and Skrondal (2006) pointed out, one set of unconditional weights is not sufficient for multilevel pseudo-maximum-likelihood estimation. Level-specific weights must be used at each level.

A number of previous researchers have found that scaling of level-1 weights could affect variance estimates (e.g., Asparouhov 2006; Pfeffermann 1993; Rabe-Hesketh & Skrondal 2006; Stapleton 2002). Several scaling methods have been explored to reduce the bias in the variance components for small cluster sizes. A common scaling method is to scale the level-1 weights to sum up to actual cluster sample size, which is the scaling method used in the simulation study of this paper. Due to space limitation, a discussion of scaling issues is not presented here. A comprehensive investigation of the weight-scaling methods could be found in the work of Asparouhov (2006).

3.4 Sandwich Estimators for Standard Errors

This section summarizes the sandwich estimator for standard errors of multilevel pseudo-maximum likelihood estimates. Detailed derivations about standard error

estimations could be found in the works of Asparouhov and Muthén (2006) and Rabe-Hesketh and Skrondal (2006).

When units are independent and identical, the standard errors can be computed using a sandwich estimator:

$$cov(\hat{\boldsymbol{\lambda}}) = \mathbf{A^{-1}BA^{-1}}, \tag{12}$$

$\mathbf{A}$ is the observed Fisher information at maximum-likelihood estimates $\hat{\boldsymbol{\lambda}}$. Let $\hat{\boldsymbol{\lambda}}'$ be the transpose of $\hat{\boldsymbol{\lambda}}$. Matrix $\mathbf{A}$ can be written as:

$$\mathbf{A} \equiv -E\left(\frac{\partial^2}{\partial\hat{\boldsymbol{\lambda}}\,\partial\hat{\boldsymbol{\lambda}}'} logL\right)\Bigg|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}}, \tag{13}$$

while $\mathbf{B}$, the outer product of the gradient vector, can be written as:

$$\mathbf{B} \equiv E\left(\frac{\partial}{\partial\hat{\boldsymbol{\lambda}}} logL\right)\left(\frac{\partial}{\partial\hat{\boldsymbol{\lambda}}'} logL\right)\Bigg|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}}. \tag{14}$$

$\mathbf{A}$ and $\mathbf{B}$ are the same when the model is the true model. In complex samples, $\mathbf{A}$ becomes the observed Fisher information at pseudo-maximum-likelihood estimates:

$$\mathbf{A}^* \equiv -E\left(\frac{\partial^2}{\partial\hat{\boldsymbol{\lambda}}\,\partial\hat{\boldsymbol{\lambda}}'} logL^*\right)\Bigg|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}}. \tag{15}$$

$\mathbf{B}$ can be obtained by summing the contributions of each independent school (Rabe-Hesketh & Skrondal 2006). Specifically, the first derivatives of the log-pseudolikelihood is:

$$\frac{\partial}{\partial\hat{\boldsymbol{\lambda}}} logL^* = \sum_{h=1}^{H}\sum_{k=1}^{K_h}\frac{\partial}{\partial\hat{\boldsymbol{\lambda}}} logL^*_{k|h}. \tag{16}$$

and $\mathbf{B}$ is calculated by:

$$\mathbf{B}^* = \sum_{h=1}^{H}\frac{K_h}{K_h-1}\sum_{k=1}^{K_h}\left(\frac{\partial}{\partial\hat{\boldsymbol{\lambda}}} logL^*_{k|h}\right)\left(\frac{\partial}{\partial\hat{\boldsymbol{\lambda}}'} logL^*_{k|h}\right). \tag{17}$$

Finally, the variances of pseudo-maximum-likelihood estimates could be estimated with:

$$cov(\hat{\boldsymbol{\lambda}}) = (\mathbf{A}^*)^{-1}\mathbf{B}^*(\mathbf{A}^*)^{-1}. \tag{18}$$

4 Simulation Design

Motivated by the versatility of the hybrid method in dealing with complex sampling weights, and by the lack of application of such technique to IRT models, this paper attempts to evaluate the performance of hybrid method in IRT models in comparison to other methods. Monte Carlo simulations are carried out to examine the performance of the above mentioned three methods in dealing with complex sample item response data.

The sample design in this paper is partly inspired by PISA 2000 sample design of the United States as described by Rabe-Hesketh and Skrondal (2006). The design includes stratification on both school and student levels, which made both the level-1 and level-2 weights informative. The simulation study chooses this design as an inspiration due to the added complexity of stratification on student level. By adopting this design, the method would be generalizable to more complex situations. Assessments with a less complex design would be a simplification of the scenario presented here.

4.1 Generating Latent Variables and Student Samples

Latent scores of the population are generated with respect to both levels (school level and student level). One level 2, the latent variable is set to follow a normal distribution with mean 0 and variance 3/7. One level 1, the latent variable is set to follow a normal distribution of mean 0 and variance 1. The setup would yield an intraclass correlation (ICC) of 0.3 for the latent variable, which is meant to mimic a fairly large clustering effect of the schools that is typically found in PISA. For example, the results from PISA2003 showed that, the ICC for math outcome was 0.345 across all countries. The ICC for USA was 0.264 (Anderson, Milford & Ross 2009). We have not found any reference on ICCs for PISA2000, but we assume them to be comparable to PISA2003. A population of 400,000 students are generated using above mentioned latent variables. The total number of schools is set to 1000 and the average school size (total number of students in a school) is set to 400. Schools are categorized into public and private schools in such a way that private schools have higher average latent scores than public schools. At level 1, students are categorized into two groups based on ethnicities. The majority group (about 70 % of all the students) are set to have a larger mean latent score than the minority group (about 30 % of the students). The proportions of minority students in each school are then identified. School type and minority status will serve as the basis for stratification in the sampling design.

The sampling method follows the design described in Sect. 2. In level-2 sampling, public schools with at least 15 % minority students are set to be twice as likely to be sampled as other schools. In level-1 sampling, minority students within public schools with at least 15 % minority students are twice as likely to be

sampled as other students. As a result, higher latent scores are associated with lower selection probabilities at both levels. Since the selection probabilities are related to the outcome measures (latent scores) on both levels, the resulting sample weights are informative on both levels. Ignoring such sampling design might lead to bias in estimations of item and person parameters in the finite population.

With this method, 75 schools are first selected in the first stage. Then 30 students are selected from each school in the second stage. The final sample has 2250 students.

4.2 Generating Item Response Data

Item response data are generated using a graded response model (Samejima 1969). The generating model is chosen for illustrative purpose only, and not meant to mimic actual PISA items. The sampled latent scores are used to generate 5-category polytomous responses for 20 items using an unidimensional graded response model, with the latent variable split into level 1 and 2. Cross-level measurement invariance is assumed. Let f^*_{ijkx} be the probability of examinee j from school k scoring x or above on item i. The model is defined as:

$$f^*_{ijkx} = f(y_{ijk} \geq x \mid \xi_{.k}, \delta_{jk}) = \frac{1}{1 + \exp(-\beta_{ix} - \alpha_i \xi_{.k} - \alpha_i \delta_{jk})}. \tag{19}$$

The examinee's probability of scoring x can be expressed as:

$$f_{ijkx} = f^*_{ijkx} - f^*_{ijk(x+1)}. \tag{20}$$

4.3 Data Analysis

The generated response data are then analyzed with four methods, namely (1) one-level modeling without weights, (2) one-level modeling with weights (design-based method), (3) two-level modeling without weights (model-based method), and (4) two-level modelling with weights at both levels (hybrid method).

In total, the simulation study has four conditions (four analytical methods). Both Mplus Version 7.2 and flexMIRT® Version 3 are used for method (1), (2) and (3). Results produced by the two packages are identical across the three methods. Only Mplus is used to conduct the analysis with method (4), as no other standard IRT packages implement the hybrid method at this moment as far as the authors are aware of. For the two multilevel models, the variance of the level-1 latent variable is set to 1, leaving the level-2 factor variance to be freely estimated. 100 replications are carried out for each condition.

5 Simulation Results

5.1 Results for Item Parameter Estimates and Standard Errors

The average item parameter estimates (slopes and intercepts) over 100 replications are plotted against the generating true values in order to gauge the biases in the point estimates. As shown in Fig. 1, the biases in point estimates are all fairly small across the four models. The point estimates in the two weighted models (right two panels) are almost unbiased. Both slope and intercept estimates are slightly upward biased in the two unweighted methods (left two panels). The weighted methods are able to yield unbiased item parameters, while the unweighted methods overestimate these point estimates.

The average estimated standard errors for slopes are plotted against the Monte Carlo standard deviations of point estimates in order to evaluate the biases in standard errors in Fig. 2. Using the Monte Carlo standard deviations as the standard, the root-mean-square errors (RMSE) of the estimated standard errors are also calculated. As we can see, the two two-level models (bottom two panels in Fig. 2) yield almost unbiased slope standard errors as the points are closely distributed around the diagonal line. The two one-level models (top two panels in Fig. 2 slightly underestimated the slope standard errors as the points are mostly under the diagonal line. The RMSE also confirm the observation.

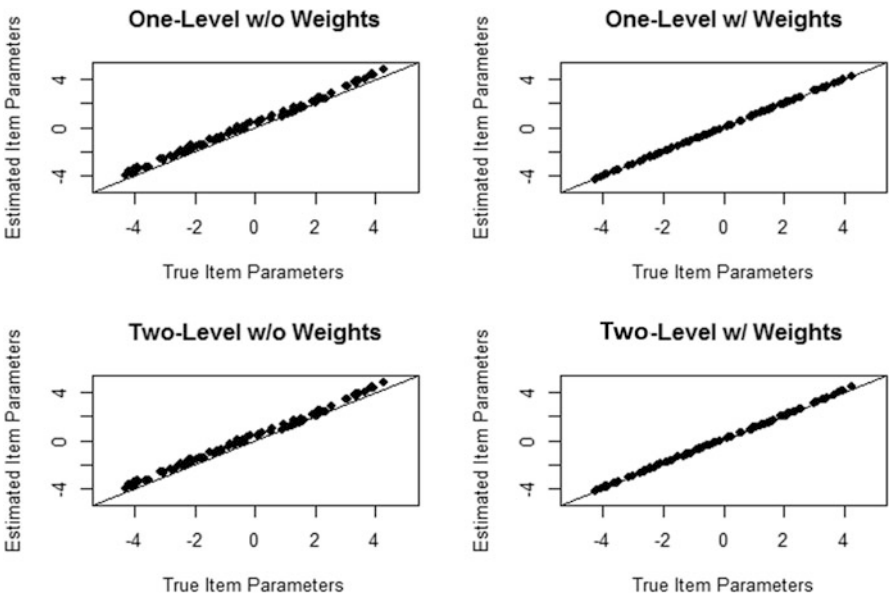


Fig. 1 True item parameters vs. estimates. The unweighted models (*left two panels*) have slightly overestimated the item parameters, while the weighted models (*right two panels*) appears to return unbiased estimates

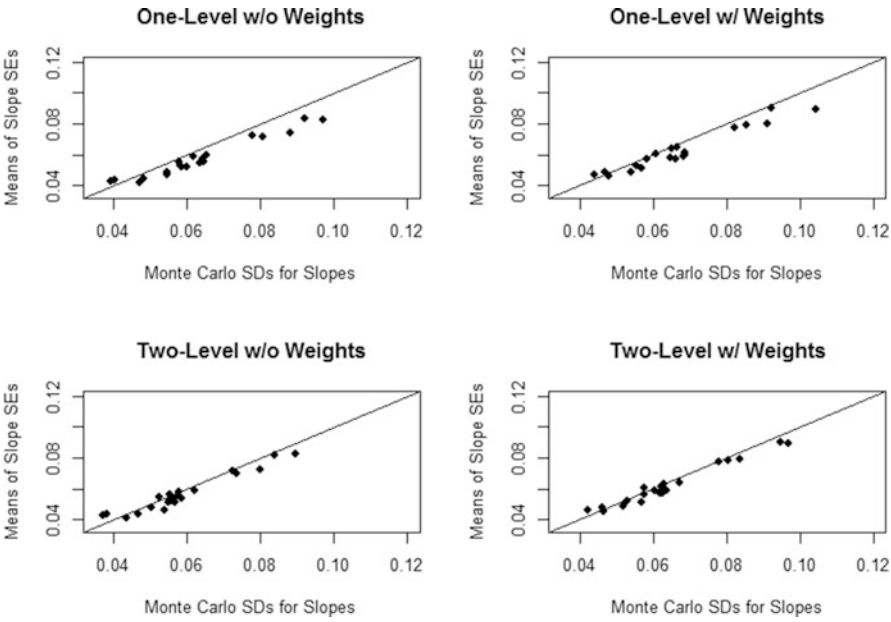


Fig. 2 Monte Carlo standard deviations for slopes vs. means of slope standard errors. The RMSEs of slope standard errors are 0.0076, 0.0065, 0.0042 and 0.0037 respectively for the one-level w/o weights, one-level w/ weights, two-level w/o weights and two-level w/ weights models

The average estimated standard errors for intercepts are plotted against the Monte Carlo standard deviations of point estimates in Fig. 3. The two two-level models (bottom two panels in Fig. 3) yield slightly biased intercept standard errors. The model-based method tends to inflate intercept standard errors, while the hybrid method slightly underestimate the intercept standard errors. The two one-level models (top two panels in Fig. 3) have severely underestimated the intercept standard errors. The RMSEs of the standard errors in the one-level models are expectedly much higher than two-level models.

The 95 % confidence intervals are constructed using the intercept estimates and their standard errors. With both the point estimate and the standard errors taken into account, the coverage rates of the true intercepts in the 95 % confidence interval are very poor in the two unweighted methods, both under 20 % across items, while the same measures for the one-level and two-level weighted methods are 86 and 90 % respectively.

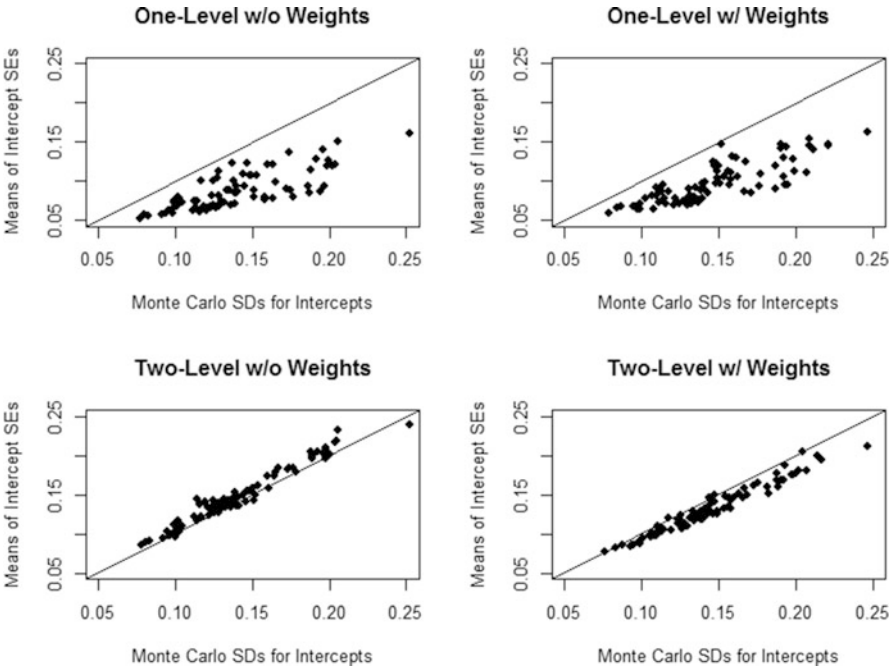


Fig. 3 Monte Carlo standard deviations for intercepts vs. means of intercept standard errors. The RMSEs of intercept standard errors are 0.0330, 0.0313, 0.0078 and 0.0129 respectively for the one-level w/o weights, one-level w/ weights, two-level w/o weights and two-level w/ weights models

5.2 Results for Level-2 Variance

Coverage of true second-level (between-school) variance in the 95 % confidence intervals of estimates is plotted in Fig. 4 for the weighted and unweighted multilevel models. It appears that the hybrid method has some advantages over traditional two-level models, as the hybrid method achieves a less biased level-2 variances and better coverage of true level-2 variance. In fact, the average percentage bias for the level-2 variance is 14 % in the unweighted model, while the same measure for the hybrid model is only −2 %. The coverage rates of true level-2 variance in the 95 % confidence intervals are 82 and 91 % respectively for the unweighted and weighted two-level models.

6 Discussion

We compared the performance of three methods to analyze item response data collected under a complex sample design, with a special interest in the performance of the pseudo-maximum-likelihood estimation method for multilevel IRT models (the hybrid method). The results show that, methods accounting for complex sample

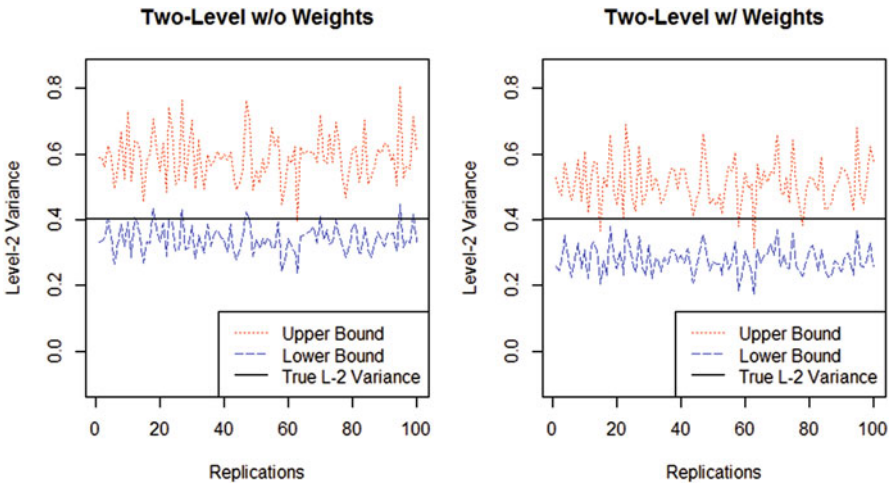


Fig. 4 Coverage of true level-2 variance in 95 % confidence intervals of estimates. The coverage rates of true level-2 variance in the 95 % confidence intervals are 82 and 91 % respectively for the unweighted and weighted two-level models

weights produce less biased point estimates for item parameters in either single-level or multilevel models, while multilevel modeling yields more accurate standard errors for item parameters than single-level models. It is worth nothing that, in the unweighted multilevel model, the coverages of the true parameters are very poor. Better standard error estimates do not seem to make up for deficiency in point estimates. The hybrid method, which accounts for both the complex sampling weights and the multilevel data structure, indeed combines the advantages of both the design-based and model-based methods. Under the unidimensional model, the performance of the hybrid method is superior to the others in terms of estimating item parameters.

The hybrid method does show great potential in analyzing testing data collected with complex sampling designs. One practical obstacle for implementing the hybrid method is the fact that it requires conditional weights for lower-level units which survey agencies generally do not release. If conditional weights are not available, and level-2 variance is not of primary interest, the authors would recommend using the total unconditional weights with single-level modeling to obtain more accurate item estimates.

There are a few limitations to the current research. First, the simulation study only uses one type of sample design. More sampling schemes should be examined to fully gauge the performance of the hybrid method, including informativeness of weights, selection mechanism, cluster size and so on. Second, the generating ICC of 0.3 in the simulation study is meant to mimic a large clustering effect. ICCs of other magnitudes should be explored to evaluate the performance of different methods. Third, an empirical illustration is missing in current research due to unavailability

of level-1 conditional weights in PISA data. Last but not least, the role of weight scaling methods has not been examined.

Our future research includes comparisons of standard errors estimated with alternative methods, evaluating the weight-scaling methods under different sample designs, and expanding the hybrid method to multi-dimensional multilevel IRT models, such as simple cluster models or testlet models.

Acknowledgements This research is supported in part by the Institute for Education Sciences, U.S. Department of Education, through grants R305D150052. The opinions expressed are those of the authors and do not represent the views of the Institute or the Department of Education. We would like to thank Dr. Li Cai for providing his unpublished paper on unidimensional model with weights and flexMIRT® program.

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*(1), 47–76.

Anderson, J. O., Milford, T., & Ross, S. P. (2009). Multilevel modeling with HLM: Taking a second look at PISA. In *Quality research in literacy and science education* (pp. 263–286). Dordrecht: Springer Netherlands.

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics - Theory and Methods, 35*(3), 439–460.

Asparouhov, T., & Muthén, B. (2006). Multilevel modelling of complex survey data. In *Proceedings of the Survey Research Methods Section, American Statistical Association 2006* (pp. 2718–2726).

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review, 51*, 279–292.

Cai, L. (2013). Multiple-group item response theory analysis in the case of complex survey data. *Contributed Paper for World Statistics Congress Session Latent Variable Modeling of Complex Survey Data*, August 2013, Hong Kong.

Fox, J., & Glas, C. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*, 269–286.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied survey data analysis*. Boca Raton, FL: CRC Press.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement, 49*(1), 82–100.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79–93.

Kish, L. (1965). *Survey sampling*. New York, NY: Wiley.

Laukaityte, I. (2013). *The importance of sampling weights in multilevel modeling of international large-scale assessment data.* Paper presented at the 9th Multilevel conference, Utrecht, March 27–29.

Mislevy, R. J., Beaton, A. E., Kaplan, B. K., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133–161.

Pfeffermann, D. (1993). The role of sampling weights when modelling survey data. *International Statistical Review, 61*, 317–337.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society. Series B, Statistical methodology, 60*, 23–40.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society, Series A, 169*(4), 805–827.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. NJ: John Wiley & Sons.

Rust, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics, 1*(4), 381–397.

Samejima, F. (1969). Estimation of Latent Ability Using a Response Pattern of Graded Scores *(Psychometric Monograph No. 17)*. Richmond, VA: Psychometric Society.

Skinner, C. J. (1989). Domain means, regression and multivariate analysis. In C. J. Skinner, D. Holt & T. M. F. Smith (Eds.), *Analysis of complex surveys* (pp. 59–87). New York, NY: Wiley.

Stapleton, L. M. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling, 9*(4), 475–502.

Scalability Coefficients for Two-Level Polytomous Item Scores: An Introduction and an Application

Daniela R. Crisan, Janneke E. van de Pol, and L. Andries van der Ark

Abstract First, we made an overview of nonparametric item response models and the corresponding scalability coefficients in Mokken scale analysis for single-level item scores and two-level dichotomous item scores. Second, we generalized these models and coefficients to two-level polytomous item scores. Third, we applied the new scalability coefficients to a real-data example, and compared the outcomes with results obtained using single-level reliability analysis and single-level Mokken scale analysis. Results suggest that coefficients from single-level analyses do not provide accurate information about scalability of two-level item scores.

Keywords Mokken scale analysis • Multilevel analysis • Nonparametric item response theory • Scalability coefficients

1 Introduction

For most tests, a single rater provides the item scores that are used to estimate a particular subject' trait value. Typically, the rater and the subject are the same person but for several clinical or pedagogical tests the rater may be, for example, the parent or the supervisor of the subject. The item scores are not nested and called single-level item scores. For some tests, multiple raters provide the item scores that are used to estimate a particular subject's trait value. Examples include teachers whose

D.R. Crisan (✉)
Department of Psychometrics and Statistics, University of Groningen, Grote Kruisstraat 2/1, 9712 TS, Groningen, The Netherlands
e-mail: d.r.crisan@rug.nl

J.E. van de Pol
Department of Education, Utrecht University, P.O. Box 80140, 3508 TC, Utrecht, The Netherlands
e-mail: j.e.vandepol@uu.nl

L.A. van der Ark
Research Institute of Child Development and Education, University of Amsterdam, P.O. Box 15776, 1001 NG, Amsterdam, The Netherlands
e-mail: l.a.vanderark@uva.nl

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_11

teaching skills are rated by all students in the classroom; hospitals for which the quality of health care is rated by multiple patients; or students whose essays are rated by multiple assessors. In these cases, the raters are nested within the subjects, and the resulting item scores are called two-level item scores.

Nonparametric item response theory (NIRT) models are flexible unidimensional item response theory (IRT) models that are characterized by item response functions that do not have a parametric form. For an introduction to NIRT models, we refer to Sijtsma and Molenaar (2002). NIRT models have been defined for dichotomous single-level item scores (Mokken 1971), polytomous single-level item scores (Molenaar 1997), and dichotomous two-level item scores (Snijders 2001), but not yet for polytomous two-level item scores.

NIRT models are attractive for two reasons. First, for single-level dichotomous item scores, NIRT models allow stochastic ordering of the latent trait by means of the unweighed sum score of the test (Grayson 1988; Hemker, Sijtsma, Molenaar & Junker 1997). This in an attractive property because for most tests the unweighed sum scores is used as a measurement value. For polytomous single-level item scores, NIRT models imply a weak form of stochastic ordering (Van der Ark & Bergsma 2010). It is unknown whether these properties carry over NIRT models for two-level item scores. Second, there are many methods available to investigate the fit of NIRT models (Mokken 1971; Sijtsma & Molenaar 2002; Van der Ark 2007). Because all well-known unidimensional item response models are a special case of the nonparametric graded response model (a NIRT model for single-level polytomous item scores) (Van der Ark 2001), investigating the fit of NIRT models is a logical first step in parametric IRT modelling: If the nonparametric graded response model does not fit, parametric IRT models will not fit either.

The set of methods to investigate the fit of NIRT models are called *Mokken scale analysis*. The most popular coefficients from Mokken scale analysis are the scalability coefficients (Mokken 1971). For a set of I items, there are $I(I-1)/2$ item-pair scalability coefficients H_{ij}, I item scalability coefficients H_i, and one total scalability coefficient H. Coefficient H reflects the accuracy of the ordering of persons using their sum scores (Mokken, Lewis & Sijtsma 1986); hence, the larger H, the more accurate is the ordering.

The remainder of this paper is organized as follows. First, we discuss NIRT models and scalability coefficients for dichotomous single-level, polytomous single-level, and dichotomous two-level item scores. Second, we generalize the NIRT model and scalability coefficients to polytomous two-level item scores, demonstrate how the scalability coefficients are estimated, and briefly discuss results from a simulation study investigating the scalability coefficients for both dichotomous and polytomous item scores (Crisan 2015). Third, we present a real-data example: We analyzed two-level polytomous item scores from the Appreciation of Support Questionnaire (Van de Pol, Volman, Oort & Beishuizen 2015), and compared the outcomes with results obtained using traditional reliability analysis. Finally, we elaborate on the implications of our findings and discuss future research directions.

2 NIRT Models and Scalability Coefficients

Let a test consists of I items, indexed by i or j. Let each item have $m + 1$ ordered response categories scored $0, \ldots, m$ indexed by x or y. If $m = 1$, the items scores are dichotomous, if $m > 1$ the item scores are polytomous. Suppose the test is used to measure the trait level of S subjects, indexed by s or t, and subject s has been rated by R_s raters, indexed by p or r. If $R_s = 1$ for all subjects, we have single-level item scores, and the index for the rater is typically omitted. Furthermore. Let X_{sri} denote the score of subject s by rater r on item i, and let X_{s++} denote the total score of subject s; that is, $X_{s++} = \sum_{i=1}^{I} \sum_{r=1}^{R_s} X_{sri}$. Finally, let θ denote a latent trait driving the item responses, and let θ_s denote the latent trait value of subject s.

2.1 NIRT Models and Scalability Coefficients for Single-Level Dichotomous Item Scores

The monotone homogeneity model (MHM) (Mokken 1971; Molenaar 1997; Sijtsma & Molenaar 2002) is a NIRT model for single-level dichotomous item scores. $P(X_{si} = x_{si}|\theta_s)$ denote the probability that subject s has score $x_{si} \in \{0, 1\}$ on item i. The MHM consists of three assumptions.

- Unidimensionality: θ is unidimensional;
- Local independence: item-scores are independent conditional on θ, that is,

$$P(X_{s1} = x_{s1}, X_{s2} = x_{s2}, \ldots, X_{sI} = x_{sI}|\theta_s) = \prod_{i=1}^{I} P(X_{si} = x_{si}|\theta_s); \qquad (1)$$

- Monotonicity: For each item i, there is a nondecreasing function $p_i(\cdot)$ such that the probability of obtaining item score 1 given latent trait value θ_s is $p_i(\theta_s) = P(X_{si} = 1|\theta_s)$.

Function $p_i(\theta)$ is known as the *item response function*. Under the MHM, item response function are allowed to intersect. If, additionally to the three assumptions, the restriction of non-intersecting of the IRFs is imposed, then the more restrictive double monotonicity model is defined (Mokken 1971).

The scalability coefficients are based on the Guttman model. Without loss of generality, let the I items be put in descending order of mean item score and be numbered accordingly, so that $P(X_i = 1) > P(X_j = 1)$ for $i < j$. The Guttman model does not allow that the easier (more popular) item has score 0 and the more difficult (less popular) item has score 1, and thus excludes item-score pattern $(X_i, X_j) = (0, 1)$, which is known as a *Guttman error*. For items i and j, let $F_{ij} = P(X_i = 0, X_j = 1)$ denote the probability of obtaining a Guttman error, and

let $E_{ij} = P(X_i = 0)P(X_j = 1)$ denote the expected probability of a Guttman error under marginal independence. Item-pair scalability coefficient H_{ij} is then defined as

$$H_{ij} = 1 - \frac{F_{ij}}{E_{ij}}. \tag{2}$$

If the MHM holds $0 \leq H_{ij} \leq 1$ for all $i \neq j$. H_{ij} equals the ratio of the covariance of X_i and X_j and the maximum covariance of X_i and X_j given the marginal item score distribution. Item scalability coefficient H_i is

$$H_i = 1 - \frac{\sum_{i \neq j} F_{ij}}{\sum_{i \neq j} E_{ij}}. \tag{3}$$

If the MHM holds $0 \leq H_i \leq 1$ for all i. H_i can be viewed as a nonparametric analogue of the discrimination parameter (Van Abswoude, Van der Ark & Sijtsma 2004). As a heuristic rule for inclusion in a scale, H_i is often required to exceed 0.3. Finally, total-scale scalability coefficient H is

$$H = 1 - \frac{\sum_i \sum_j F_{ij}}{\sum_i \sum_j E_{ij}}. \tag{4}$$

As a heuristic rule, $0.3 < H \leq 0.4$ is considered a weak scale, $0.4 < H \leq 0.5$ is considered a moderate scale, and $H > 0.4$ is considered a strong scale.

2.2 NIRT Models and Scalability Coefficients for Single-Level Polytomous Item Scores

The nonparametric graded response model (a.k.a. the MHM for polytomous items (Molenaar 1997) is the least restrictive NIRT model for polytomous items. As the MHM, it consists of the assumptions unidimensionality, local independence, and monotonicity but monotonicity is defined differently. For item score x ($x = 1, \ldots, m$) for each item i there is a nondecreasing function $p_{ix}(\cdot)$ such that the probability of obtaining at least item score x given latent trait value θ_s is $p_{ix}(\theta_s) = P(X_{si} \geq x|\theta_s)$. Function $p_{ix}(\theta)$ is known as the *item step response function*. Under the nonparametric graded response model, ISRFs from the same item cannot intersect by definition but ISRFs from different items are allowed to intersect. If, additionally to the three assumptions the restriction of non-intersecting of the ISRFs is imposed, then we have the more restrictive double monotonicity model for polytomous items (Molenaar 1997).

Scalability coefficients for polytomous item scores are more complicated than for dichotomous item scores, which are a special case. They are best explained using an

Table 1 Frequency table for two polytomous items with three response categories

		Item 2			
	Response	0	1	2	$P(X_1 \geq x)$
Item 1	0	**2** (0)	1 (2)	0 (4)	1
	1	**3** (0)	0 (1)	0 (2)	3/4
	2	**3** (0)	**2** (0)	**1** (0)	1/2
$P(X_2 \geq x)$		1	1/3	1/12	

Note: Frequencies not pertaining to Guttman errors are in boldface, frequencies pertaining to Guttman errors are in normal font, Guttman weights are between parentheses. The last row and column show the marginal cumulative probabilities

example. Table 1 contains the scores of 12 subjects on two items, each having three ordered answer categories.

First, Guttman errors are determined. *Item steps* (Molenaar 1983) $X_i \geq x (i = 1, \ldots, I; x = 1, \ldots, m)$ are boolean expressions indicating whether or not an item score is at least x. $P(X_i \geq x)$ defines the popularity of item step $X_i \geq x$. The item steps are placed in descending order of popularity. For the data in Table 1, the order of the item-steps is:

$$X_1 \geq 1, X_1 \geq 2, X_2 \geq 1, X_2 \geq 2. \tag{5}$$

Items steps $X_1 \geq 0$ and $X_2 \geq 0$ are omitted because, by definition, $P(X_1 \geq 0) = P(X_2 \geq 0) = 1$. Item-score pattern (x, y) is a Guttman error if an item step that has been passed is preceded by an item step that has not been passed. Let z_g^{xy} indicate whether (score 1) or not (score 0) the gth ordered item step has been passed for item-score pattern (x, y). The values of z_g^{xy} are collected in vector $\mathbf{z}^{xy} = (z_1^{xy}, \ldots, z_G^{xy})$. To obtain item-score pattern $(0, 2)$ in Table 1, a subject must have passed item steps $X_2 \geq 1$ and $X_2 \geq 2$ but not item steps $X_1 \geq 1$ and $X_1 \geq 2$. Hence, for item-score pattern $(0, 2)$, $\mathbf{z}^{02} = (0, 0, 1, 1)$. Because item steps that have been passed are preceded by items steps that have not been passed, $(0, 2)$ is identified as a Guttman error. Similarly, for item-score pattern $(2, 1)$, $\mathbf{z}^{21} = (1, 1, 1, 0)$ and item-score pattern $(2, 1)$ is not a Guttman error. In Table 1, the four item-score patterns for which the frequencies are printed in normal font are Guttman errors, whereas the frequencies printed in bold font are not.

Second, the frequencies of the item-score patterns are weighed (Molenaar 1991); the weight being equal to the number of times an item step that has not been passed preceded an item step that has been passed. Weight w_{ij}^{xy} equals

$$w_{ij}^{xy} = \sum_{h=2}^{G} \left\{ z_h^{xy} \times \left[\sum_{g=1}^{h-1} (1 - z_g^{xy}) \right] \right\} \tag{6}$$

(Kuijpers, Van der Ark & Croon 2013; Ligtvoet, Van der Ark, te Marvelde & Sijtsma 2010). For example, for item-score pattern $(0, 2)$, $\mathbf{z}^{02} = (z_1^{02}, z_2^{02}, z_3^{02}, z_4^{02}) = (0, 0, 1, 1)$. Using Eq. (6), the weight equals $w_{ij}^{02} = 4$. Table 1 shows the weights between parentheses.

Item-pair scalability coefficient H_{ij} for polytomous items is

$$H_{ij} = 1 - \frac{\sum_x \sum_y w_{ij}^{xy} P(X_i = x, X_j = y)}{\sum_x \sum_y w_{ij}^{xy} P(X_i = x) P(X_j = y)} \tag{7}$$

(Molenaar 1991). Because item-score patterns that are not Guttman errors have weight 0, the probabilities pertaining to these patterns do not count, and the numerator of Eq. (7) is simply the sum of observed weighed Guttman errors, and the denominator the sum of expected weighed Guttman errors. Similarly, item scalability coefficient H_i for polytomous items is

$$H_j = 1 - \frac{\sum_{i \neq j} \sum_x \sum_y w_{ij}^{xy} P(X_i = x, X_j = y)}{\sum_{i \neq j} \sum_x \sum_y w_{ij}^{xy} P(X_i = x) P(X_j = y)}, \tag{8}$$

and the total scale scalability coefficient H is

$$H = 1 - \frac{\sum \sum_{i \neq j} \sum_x \sum_y w_{ij}^{xy} P(X_i = x, X_j = y)}{\sum \sum_{i \neq j} \sum_x \sum_y w_{ij}^{xy} P(X_i = x) P(X_j = y)}. \tag{9}$$

Note that for dichotomous items, the Guttman error receives a weight 1, and Eqs. (7)–(9) reduce to Eqs. (2)–(4), respectively. In Table 1, because there are only two items, $H_{12} = H_1 = H_2 = H = 0.50$.

2.3 NIRT Models and Scalability Coefficients for Two-Level Dichotomous Item Scores

Snijders (2001) generalized the MHM for dichotomous items to two-level data. As in the MHM, each subject has a latent trait value θ_s. In addition, rater r is assumed to have a deviation (δ_{sr}), so the latent trait value for subject s as rated by rater r is $\theta_s + \delta_{sr}$. Deviation δ_{sr} can be considered a random rater effect together with the subject by rater interaction. It is assumed that the raters are a random sample from the population of raters, so deviations δ_{sr} can be considered independent and randomly distributed variables. As the MHM, Snijders' model for two-level data assumes unidimensionality, local independence, and monotonicity for the item response functions $p_i(\theta_s + \delta_{sr}) = P(X_{sri} = 1|\theta_s; \delta_{sr})$. In addition, a second nondecreasing item response function is defined $\pi_i(\theta_s) = P(X_{si} = 1|\theta_s) = E_\delta[p_i(\theta_s + \delta_{sr})]$. If $p_i(\theta_s + \delta_{sr})$ is nondecreasing, then so is $\pi_i(\theta_s)$, yet $\pi_i(\theta_s)$ will be flatter.

Snijders generalized scalability coefficients for dichotomous items [Eqs. (2)–(4)] to two-level data, resulting in *within-rater* and *between-rater* scalability coefficients.[1] The within-rater scalability coefficients H_{ij}^W, H_i^W, and H^W are in fact equivalent to the scalability coefficients that were defined for the MHM [Eqs. (2)–(4), respectively], where every rater-subject combination is considered a separate case.

Snijders defined the between-rater item-pair scalability coefficients

$$H_{ij}^B = 1 - \frac{P(X_{sri} = 1, X_{spj} = 0)}{P(X_{sri} = 1)P(X_{srj} = 0)} (p \neq r). \tag{10}$$

The joint probability in the numerator is computed for pairs of different raters p and r ($p \neq r$) nested within the same subject s. More specifically, the numerator represents the joint probability that rater r assigns score 1 on item i to subject s and rater p assigns score 0 on item j to subject s. Because the denominator consists of a product of two probabilities that are independent of r, replacing r with p in the second term of the denominator would not make any difference: the expected proportion of Guttman errors under marginal independence remains the same. Using a similar line of reasoning, the item between-rater scalability coefficients are

$$H_i^B = 1 - \frac{\sum_{j \neq i} P(X_{sri} = 1, X_{spj} = 0)}{\sum_{j \neq i} P(X_{sri} = 1)P(X_{srj} = 0)} (p \neq r) \tag{11}$$

and

$$H^B = 1 - \frac{\sum \sum_{j \neq i} P(X_{sri} = 1, X_{spj} = 0)}{\sum \sum_{j \neq i} P(X_{sri} = 1)P(X_{srj} = 0)} (p \neq r). \tag{12}$$

Within-rater scalability coefficients are useful for investigating the quality of the test as a unidimensional cumulative scale for subject-rater combinations. The between-rater scalability coefficients and the ratio of the within- and between-rater scalability coefficients are useful for investigating the extent to which item responses are driven by the subjects trait value rather than by rater effects. If Snijders' model holds, $0 < H^B \leq H^W$ (Snijders 2001); and larger values indicate greater scalability. In the extreme case that there is no rater variation ($\delta_{rs} = 0$ for all r and all s), $H^B = H^W$. As a heuristic rule, Snijders suggested $H^B > 0.1$ and $H^W > 0.2$ to be reasonable. The ratio of the two scalability coefficients reflect the relative effect of the subjects and the raters. Low values indicate that the effect of raters is large and many raters per subject are required to scale the subjects. Snijders suggested $H^B/H^W \geq 0.3$ could be labelled reasonable and $H^B/H^W \geq 0.6$ excellent. The measurement for scaling subjects is the mean total score of a subjects across all raters: $\overline{X}_{s++}$.

[1] Terminology is ours; Snijders used within-subject and between-subject scalability.

3 A Generalization to Two-Level Polytomous Item Scores

Given the work on scalability coefficients for single-level polytomous item scores (Sect. 2.2) and two-level dichotomous item scores (Sect. 2.3), a generalization to two-level polytomous item scores is rather straightforward. The within-rater scalability coefficients for polytomous item scores are the same as the scalability coefficients for single-level polytomous item scores [Eqs. (7)–(9)] when considering all rater-subjects combinations as individual cases.

The between-rater scalability coefficients are defined as follows:

$$H_{ij}^{B} = 1 - \frac{\sum_x \sum_y w_{ij}^{xy} P(X_{sri} = x, X_{spj} = y)}{\sum_x \sum_y w_{ij}^{xy} P(X_{sri} = x) P(X_{srj} = y)} (p \neq r), \tag{13}$$

$$H_{i}^{B} = 1 - \frac{\sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} P(X_{sri} = x, X_{spj} = y)}{\sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} P(X_{sri} = x) P(X_{srj} = y)} (p \neq r), \tag{14}$$

and

$$H^{B} = 1 - \frac{\sum \sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} P(X_{sri} = x, X_{spj} = y)}{\sum \sum_{j \neq i} \sum_x \sum_y w_{ij}^{xy} P(X_{sri} = x) P(X_{srj} = y)} (p \neq r). \tag{15}$$

It may be verified that in case of dichotomous item scores Eqs. (13)–(15) reduce to Equations reduces to (10)–(12), respectively.

3.1 Estimation of the Scalability Coefficients

Snijders (2001) proposed estimators for the scalability coefficients for dichotomous item scores, by substituting the probabilities in their defining formulas by relative frequencies. If the number of raters per subject (R_s) is not the same for all subjects, then the probabilities required to compute the scalability coefficients can be estimated by averaging the relative frequencies across subjects. Snijders' estimators can be generalized to polytomous item scores. Let $\mathbf{1}(X_{sri} = x)$ denote the indicator function that $X_{sri} = x$, and let $\widehat{P}_i(x)$ be the estimator for $P(X_{sri} = x)$; then,

$$\widehat{P}_i(x) = \frac{1}{S} \sum_s \frac{1}{R_s} \sum_r \mathbf{1}(X_{sri} = x). \tag{16}$$

Equation (16) determines the proportions of raters per subject with a score x on item i and then averages these proportions across subjects, yielding the estimated probability of a score equal to x on item i.

The joint probabilities in the numerators of the scalability coefficients can be estimated as follows. Let $\widehat{P}_{ij}^{W}(x, y)$ denote the estimated within-rater joint probability

that $X_{sri} = x$ and $X_{srj} = y$, and let $\widehat{P}^B_{ij}(x, y)$ denote the estimated between-rater joint probability that $X_{sri} = x$ and $X_{spj} = y$. Then,

$$\widehat{P}^W_{ij}(x, y) = \frac{1}{S} \sum_s \frac{1}{R_s} \sum_r \mathbf{1}(X_{sri} = x, X_{srj} = y), \tag{17}$$

and

$$\widehat{P}^B_{ij}(x, y) = \frac{1}{S} \sum_s \frac{1}{R_s(R_s - 1)} \sum\sum_{p \neq r} \mathbf{1}(X_{sri} = x, X_{spj} = y). \tag{18}$$

Finally, substituting the probabilities in the defining formulas of the scalability coefficients with the estimators in Eqs. (16)–(18) leads to the following estimators of the within- and between-subject scalability coefficients:

$$\widehat{H}^W_{ij} = 1 - \frac{\sum_x \sum_y w^{xy}_{ij} \widehat{P}^W_{ij}(x, y)}{\sum_x \sum_y w^{xy}_{ij} \widehat{P}_i(x) \widehat{P}_j(y)}, \tag{19}$$

$$\widehat{H}^B_{ij} = 1 - \frac{\sum_x \sum_y w^{xy}_{ij} \widehat{P}^B_{ij}(x, y)}{\sum_x \sum_y w^{xy}_{ij} \widehat{P}_i(x) \widehat{P}_j(y)}, \tag{20}$$

$$\widehat{H}^W_{i} = 1 - \frac{\sum_{j \neq i} \sum_x \sum_y w^{xy}_{ij} \widehat{P}^W_{ij}(x, y)}{\sum_{j \neq i} \sum_x \sum_y w^{xy}_{ij} \widehat{P}_i(x) \widehat{P}_j(y)}, \tag{21}$$

$$\widehat{H}^B_{i} = 1 - \frac{\sum_{j \neq i} \sum_x \sum_y w^{xy}_{ij} \widehat{P}^B_{ij}(x, y)}{\sum_{j \neq i} \sum_x \sum_y w^{xy}_{ij} \widehat{P}_i(x) \widehat{P}_j(y)}, \tag{22}$$

$$\widehat{H}^W = 1 - \frac{\sum \sum_{j \neq i} \sum_x \sum_y w^{xy}_{ij} \widehat{P}^W_{ij}(x, y)}{\sum \sum_{j \neq i} \sum_x \sum_y w^{xy}_{ij} \widehat{P}_i(x) \widehat{P}_j(y)}, \tag{23}$$

and

$$\widehat{H}^B = 1 - \frac{\sum \sum_{j \neq i} \sum_x \sum_y w^{xy}_{ij} \widehat{P}^B_{ij}(x, y)}{\sum \sum_{j \neq i} \sum_x \sum_y w^{xy}_{ij} \widehat{P}_i(x) \widehat{P}_j(y)}. \tag{24}$$

Example 1 illustrates the computation of the scalability coefficients.

Example 1. Table 2 (upper panel) shows the frequencies of the scores on 2 items, each having 3 ordered response categories, assigned by 12 raters to 3 subjects: Four raters rated subject 1 ($R_1 = 4$), two raters rated subject 2 ($R_2 = 3$), and five raters rated subject 3 ($R_3 = 5$). Frequencies equal to zero are omitted. These frequencies equal $\sum_r \mathbf{1}(X_{sri} = x, X_{srj} = y)$ and are required for computing $\widehat{P}^W_{ij}(x, y)$ (Eq. (17); values in last row of Table 2, upper panel). For example, $\hat{P}^W_{12}(0, 0) = \frac{1}{3}(\frac{1}{4} \times 2 + 0 + 0) \approx 0.17$.

Table 2 Frequencies of observed item-score patterns per subject (upper panel), frequencies of observed item-score patterns where each item-score in a pattern is assigned by a different rater for each subject (middle panel), and marginal frequencies of observed item-score patterns per subject (lower panel)

	Item-score pattern (x, y)									
s	(0,0)	(0,1)	(0,2)	(1,0)	(1,1)	(1,2)	(2,0)	(2,1)	(2,2)	R_s
1	2	1		1						4
2				1				2		3
3				1			3		1	5
$\hat{P}^W_{12}(x, y)$	0.17	0.08	0.00	0.26	0.00	0.00	0.20	0.22	0.07	

	Item-score pattern (x, y)									
s	(0,0)	(0,1)	(0,2)	(1,0)	(1,1)	(1,2)	(2,0)	(2,1)	(2,2)	$R_s(R_s - 1)$
1	7	2		2		2				12
2							2	2		6
3				3		1	13		3	20
$\hat{P}^B_{12}(x, y)$	0.19	0.06	0.00	0.11	0.00	0.07	0.33	0.11	0.05	

	Item 1			Item 2			
s	$x = 0$	$x = 1$	$x = 2$	$x = 0$	$x = 1$	$x = 2$	R_s
1	3	1		3	1		4
2		1	2	1	2		3
3		1	4	4		1	5
$\hat{P}_i(x)$	0.25	0.26	0.49	0.63	0.31	0.07	

Note: unobserved item-score patterns are left blank

Table 2 (middle panel) shows the frequencies of the item-score patterns assigned by different raters (e.g., $\sum\sum_{p\neq r} \mathbf{1}(X_{sri} = x, X_{spj} = y)$). For example, score 7 (first row, first column) is obtained as follows. Subject 1 received four item-score patterns: (0,0); (0,0); (0,1); and (1,0). Within these four patterns, it occurs 7 times that one rater has score 0 on item 1 and a different rater has score 0 on item 2. Then, $\hat{P}^B_{12}(0, 0) = \frac{1}{3}(\frac{1}{12} \times 7 + 0 + 0) \approx 0.19$.

Table 2 (lower panel) shows he marginal frequencies of the item scores for each subject (i.e., $\sum_r \mathbf{1}(X_{sri} = x)$), required for estimating $\widehat{P}_i(x)$ [Eq. (16)]. For example, $\hat{P}_1(0) = \frac{1}{3} \times (\frac{1}{4} \times 3 + 0 + 0) = 0.25$. Using the weights from Table 1 yields $\hat{H}^W_{12} = \widehat{H}^W_1 = \widehat{H}^W_2 = \widehat{H}^W = 0.50$, and $\hat{H}^B_{12} = \widehat{H}^B_1 = \widehat{H}^B_2 = \widehat{H}^B = 0.15$.

3.2 Results from a Simulation Study

Crisan (2015) performed a simulation study to the effect of item discrimination, number of ordered answer categories, the variance ratio of θ and δ, the number of subjects, and the number of raters per subject on the magnitude of $\hat{H}_W$, $\hat{H}_B$, and the ratio of $\hat{H}_B$ and $\hat{H}_W$. We briefly reiterate the main results here.

The variance ratio of θ and δ had an extremely large positive effect on the magnitude of $\widehat{H}_B$ ($\eta^2 = 0.985$) and $\hat{H}_B/\hat{H}_W$ ($\eta^2 = 0.558$), whereas item discrimination had an extremely large positive effects on the magnitude $\hat{H}_W$ ($\eta^2 = 0.766$) and $\hat{H}_B$ ($\eta^2 = 0.280$). Finally number of ordered answer categories had a very large positive effect of the magnitude of $\widehat{H}_W$. The variance ratio of θ and δ and number of subjects had the largest effects on the precision of the estimated values of $\hat{H}_W$, $\hat{H}_B$, and $\hat{H}_B/\hat{H}_W$.

4 Real-Data Example

We analyzed item scores of the Appreciation of Support Questionnaire (ASQ) (Van de Pol et al. 2015). The ASQ consists of 11 polytomously scored items (Translated items in Table 3). For each item, the scores ranged from 0 ("I don't agree at all") to 4 ("I totally agree"). The data came from an experimental study on the effects of scaffolding on prevocational students' achievement, task effort, and appreciation of support (Van de Pol et al. 2015). Six hundred fifty nine grade-8 students in The Netherlands, nested in 30 teachers, used the ASQ to express their appreciation of their own teacher's support. The number of students per teacher ranged from 12 to 46 ($M = 21.97$, $SD = 5.91$).

We conducted traditional reliability analysis, traditional Mokken scale analysis, and two-level Mokken scale analysis. Traditional reliability analysis and traditional Mokken scale analysis are inappropriate analyses for these data. However, they

Table 3 The items if the appreciation of support questionnaire

Item	Content	*M*	*SD*	IRC
1	The advice that this teacher gave me and my group was very helpful	2.53	1.00	0.70
2	Because of the way in which this teacher helped me and my group, I could focus on my work with ease	2.24	1.02	0.67
3	I felt the teacher took me seriously because of the way he/she helped me and my group	2.75	0.97	0.61
4	Because of the way this teacher helped me and my group, I could really learn new things	2.37	1.03	0.71
5	Because of the way this teacher helped me and my group, I made an effort	2.42	0.93	0.71
6	The way in which this teacher helped me and my group really worked for me	2.22	0.98	0.72
7	I could really use the help that this teacher offered	2.49	1.01	0.75
8	I worked hard with this teacher	2.37	0.98	0.67
9	The way in which this teacher helped me and my group was pleasant	2.46	1.03	0.77
10	The explanation and help of this teacher was really helpful	2.39	0.99	0.77
11	Because of the explanation and help of this teacher, I could proceed	2.48	1.03	0.71

Note: M = Mean, SD = standard deviation, IRC = item rest correlation

are used to demonstrate the different outcomes. All analyses were conducted in `R` (R Core Team 2015) using the packages `psych` (Revelle 2015) and `CTT` (Willse 2014) for traditional reliability analysis, `mokken` (Van der Ark 2007) for one-level Mokken scale analysis, and code available from the first author for two-level Mokken scale analysis.

4.1 Reliability Analysis

In traditional reliability analysis the nested structure is ignored. The descriptive statistics of the item scores were all similar: mean item scores ranged between 2.22 and 2.75, the item standard deviations ranged between 0.97 and 1.03, and the item rest correlations ranged between 0.61 and 0.75 (Table 3). Cronbach's alpha was 0.93. These results suggest a very reliable test score with no indication that items should be revised. The test score had mean $M = 26.72$, standard deviation $SD = 8.41$.

4.2 One-Level Mokken Scale Analysis

In one-level Mokken scale analysis, the nested structure is also ignored. Table 4 shows the item-pair and item scalability coefficients plus standard errors (Kuijpers et al. 2013). Because all item-pair scalability coefficients were greater than 0, and all item scalability coefficients are greater than default lower bound $c = 0.3$, the 11 items form a Mokken scale. The total scalability coefficient equalled $H = 0.58(0.02)$, which qualifies as a strong scale. In addition, we investigated monotonicity using the method *manifest monotonicity* (Junker & Sijtsma 2000), local independence using Ellis' theoretical upper and lower bounds (Ellis 2014), and non-intersection using the method *pmatrix* (Mokken 1971). We found no evidence of any substantial violation of the MHM and the double monotonicity model.

4.3 Two-Level Mokken Scale Analysis

From the single-level Mokken scale analysis we concluded that the assumptions of the double monotonicity model are reasonable. The within-rater scalability coefficients are the same as the scalability coefficients in single-level Mokken scale analysis (Table 4). The between-rater scalability coefficients (Table 5; upper diagonal and penultimate row) are greater than Snijder's heuristic lower bound 0.1 suggesting a satisfactory consistency between the raters. The total-scale between-rater scalability coefficient equalled $H^B = 0.14$. The ratio of the between and

Table 4 Scalability coefficients and standard errors for the appreciation of support questionnaire

	Item										
Item	1	2	3	4	5	6	7	8	9	10	11
1		0.60	0.55	0.60	0.50	0.58	0.64	0.47	0.58	0.60	0.57
2	0.04		0.49	0.53	0.62	0.52	0.55	0.60	0.58	0.57	0.50
3	0.04	0.04		0.53	0.51	0.54	0.56	0.53	0.58	0.52	0.52
4	0.04	0.04	0.04		0.57	0.60	0.57	0.52	0.60	0.60	0.54
5	0.04	0.03	0.04	0.03		0.64	0.60	0.67	0.62	0.59	0.53
6	0.04	0.04	0.04	0.03	0.03		0.61	0.58	0.70	0.68	0.57
7	0.03	0.04	0.04	0.04	0.03	0.03		0.54	0.67	0.63	0.67
8	0.04	0.03	0.04	0.04	0.03	0.03	0.04		0.57	0.56	0.50
9	0.04	0.04	0.04	0.03	0.03	0.03	0.03	0.03		0.68	0.60
10	0.04	0.03	0.04	0.03	0.03	0.03	0.04	0.03	0.03		0.67
11	0.03	0.04	0.04	0.04	0.04	0.04	0.03	0.04	0.03	0.03	
H_i	0.57	0.56	0.53	0.57	0.58	0.60	0.60	0.55	0.62	0.61	0.57
SE	0.02	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.03

Note: item-pair scalability coefficients H_{ij} are in the upper-triangular matrix, the standard errors in the lower-triangular matrix. Item scalability coefficients H_i and standard errors are in the last two rows

Table 5 Between-subject H coefficients for the appreciation of support questionnaire

	Item										
Item	1	2	3	4	5	6	7	8	9	10	11
1		0.16	0.13	0.17	0.15	0.17	0.15	0.18	0.16	0.16	0.13
2	0.27		0.11	0.14	0.15	0.13	0.13	0.15	0.15	0.14	0.12
3	0.23	0.23		0.13	0.12	0.11	0.10	0.12	0.11	0.11	0.09
4	0.27	0.25	0.24		0.15	0.14	0.14	0.16	0.15	0.15	0.13
5	0.30	0.25	0.24	0.25		0.14	0.12	0.16	0.15	0.12	0.11
6	0.29	0.24	0.20	0.23	0.22		0.14	0.16	0.14	0.16	0.12
7	0.23	0.24	0.18	0.24	0.21	0.23		0.15	0.14	0.14	0.12
8	0.39	0.25	0.22	0.31	0.24	0.28	0.28		0.17	0.15	0.13
9	0.27	0.26	0.19	0.25	0.24	0.20	0.21	0.30		0.15	0.13
10	0.27	0.24	0.21	0.25	0.21	0.23	0.22	0.27	0.22		0.13
11	0.23	0.24	0.18	0.24	0.21	0.21	0.18	0.26	0.21	0.19	
H_i^B	0.16	0.14	0.11	0.14	0.14	0.14	0.13	0.15	0.15	0.14	0.12
H_i^B/H_i^W	0.27	0.25	0.21	0.25	0.23	0.23	0.22	0.28	0.23	0.23	0.21

Note: item-pair scalability coefficients H_{ij}^B are in the upper-triangular matrix, the ratio of H_{ij}^B and H_{ij}^W in the lower-triangular matrix. Item scalability coefficients H_i^W and H_i^B/H_i^W are in the last two rows

within scalability coefficients (lower diagonal and last row) ranged from 0.18–0.27. All values are less than 0.3, (Snijder's heuristic value of a reasonable scale). This suggests that the rater deviation is relatively large and more students may be required for the scaling of these teachers. The results from the two-level scaling analysis shows a less bright picture than the results from the one-level analyses. Finally, the mean and standard deviation of the subject scores $\overline{X}_s$ were $M = 26.8$ and $SD = 4.35$, respectively.

5 Discussion

This chapter presented a first step in reviving Mokken scale analysis for two-level data, a method that has been largely ignored since its introduction 15 years ago. Our main contribution is the generalization of Snijder's (Snijders 2001) scalability coefficients to polytomous items. We have some reservations because the scalability coefficients for two-level polytomous data were derived by analogy, and without formal proof that the properties of the scalability coefficients for two-level polytomous item scores behave as one would expect under a two-level polytomous NIRT model.

Furthermore, using guidelines from Snijders (2001) and Crisan (2015) in the analysis of a real-data example, we showed that ignoring the two-level structure may result in at least two problems: First, single-level analyses provide information about the raters' scores rather than the subjects scores, whereas the interest is in scaling subjects, not raters. This problem has not always been acknowledged. Second, interpreting the quality of the scale using single-level statistics may give an that is too optimistic. Therefore, it is important that Mokken scale analysis for two-level data is developed further. A possible next step is the derivation of standard errors for the scalability coefficients proposed in this paper. If that has been accomplished the bias and variance of both the point estimates and standard errors can be investigated. Second, it would be interesting to investigate whether other methods in Mokken scale analysis can be generalized to multi-level data. As a start, Snijders proposed using the intra-subject correlation coefficient to assess reliability in two-level item scores, which has been generalized to polytomous items by Crisan (2015). Finally, the current methods should be further extended so that a rater is allowed to assess multiple subjects, and the methods should be implemented in software; both would increase the range of possible applications.

Acknowledgements We would to thank Letty Koopman for commenting on the first draft of the paper.

References

Crisan, D. R. (2015). *Scalability coefficients for two-level dichotomous and polytomous data: A simulation study and an application* Unpublished master's thesis. Tilburg University, Tilburg.

Ellis, J. L. (2014). An inequality for correlations in unidimensional monotone latent variable models for binary variables. *Psychometrika, 79*, 303–316.

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika, 53*, 383–392.

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*, 331–217.

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24*, 65–81.

Kuijpers, R. E., Van der Ark, L. A., & Croon, M. A. (2013). Standard errors and confidence intervals for scalability coefficients in Mokken scale analysis using marginal models. *Sociological Methodology, 43*, 42–69.

Ligtvoet, R., Van der Ark, L. A., te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement, 70*, 578–595.

Mokken, R. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton.

Mokken, R., Lewis, C., & Sijtsma, K. (1986). Rejoinder to "The Mokken scale: A critical discussion". *Applied Psychological Measurement, 10*, 279–285.

Molenaar, I. W. (1983). *Item steps*. (Heymans Bulletins HB-83-630-EX). Groningen: University of Groningen.

Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden, 12*(37), 97–117.

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369–380). New York, NY: Springer.

R Core Team (2015). R: A language and environment for statistical computing [computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

Revelle, W. (2015). Psych: Procedures for personality and psychological research [computer software]. Evanston, IL: Northwestern University. Retrieved from http://CRAN.R-project.org/package=psychVersion=1.5.8.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Snijders, T. A. B. (2001). Two-level nonparametric scaling for dichotomous data. In A. Boomsma, M. A. J. van Duijn, & T. A B. Snijders (Eds.), *Essays on item response theory* (pp. 319–338). New York, NY: Springer.

Van Abswoude, A. A. H., Van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement, 28*, 3–24.

Van de Pol, J., Volman, M., Oort, F., & Beishuizen, J. (2015). The effects of scaffolding in the classroom: support contingency and student independent working time in relation to student achievement, task effort and appreciation of support. *Instructional Science, 43*, 615–641.

Van der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement, 25*, 273–282.

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*(11), 1–19.

Van der Ark, L. A., & Bergsma, W. P. (2010). A note on stochastic ordering of the latent trait using the sum of polytomous item scores. *Psychometrika, 75*, 272–279.

Willse, J. T. (2014). CTT: Classical test theory functions. R package version 2.1 [computer software]. Retrieved from http://CRAN.R-project.org/package=CTT.

Numerical Differences Between Guttman's Reliability Coefficients and the GLB

Pieter R. Oosterwijk, L. Andries van der Ark, and Klaas Sijtsma

Abstract For samples smaller than 1000 observations and tests longer than ten items, the greatest lower bound (GLB) to the reliability is known to be biased and not recommended as a method to estimate test-score reliability. As a first step in finding alternative lower bounds under these conditions, we investigated the population values of seven reliability coefficients: Coefficients λ_1, λ_2, λ_3 (a.k.a Cronbach's alpha), λ_4, λ_5, λ_6 and the GLB under varying correlational structures, and varying levels of number of items and item variances. Coefficients λ_2, λ_4 and λ_6 had population values closest to the GLB and may be considered as alternatives for the GLB in small samples. A necessary second step, investigating the behavior of these coefficients in samples, is a topic for future research.

Keywords Classical test theory • Greatest lower bound • Guttman's lambda coefficients • Reliability

1 Introduction

The purpose of this study was to compare seven methods for computing a test score's reliability. The methods are reliability coefficients proposed by Guttman (1945) and the greatest lower bound to the reliability (GLB; Bentler & Woodward 1980; Ten Berge, Snijders, & Zegers 1981; Woodhouse & Jackson 1977). Guttman's coefficients are known as λ_1 through λ_6, of which λ_3 equals the well known coefficient α (e.g., Cronbach 1951). Results in this study were obtained for the population (i.e., parameters).

In the population, the GLB is known to be closest to the reliability (Sijtsma 2009) of all lower bounds used in classical test theory (CTT). In samples, due to chance capitalization, the GLB is known to overestimate test-score reliability when

P.R. Oosterwijk (✉) • K. Sijtsma
Tilburg University, Tilburg, The Netherlands
e-mail: p.r.oosterwijk@gmail.com; K.Sijtsma@tilburguniversity.edu

L.A. van der Ark
University of Amsterdam, Amsterdam, The Netherlands
e-mail: L.A.vanderArk@uva.nl

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer
Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_12

the sample size is smaller than 1000 and the test length exceeds ten items (Ten Berge & Sočan 2004). The question at hand is whether the λ coefficients should be recommended as alternative lower bounds under these conditions. A first step, in answering this question is to investigate whether in the population, the values of the λ coefficients are close enough to the GLB to be viable candidates. This step is investigated in this paper. A second step is to investigate the bias and variance of the λ coefficients in samples. In particular, λ_4, λ_5, λ_6 are the result of maximization procedures and may be prone to chance capitalization just like the GLB. This second step is currently being investigated by the authors.

It is known that λ_1 is smaller than λ_3, that λ_3 is smaller than λ_2, and that all three coefficients are smaller than the GLB, but the relationships of the other three λs with λ_1, λ_2, and λ_3, with the GLB, and with one another are either unknown or only known for special situations. Also, for most λs it is unknown how much their values differ from each other, and how much they differ from the GLB. This study discusses the mutual relationships between the seven methods at the theoretical level, and uses a computational study to focus on the issue of numerical differences between the seven coefficients.

In addressing the numerical differences between the λs and the GLB, we assumed that differences varied across different test and item properties. We investigated the influence of the following factors on the values of seven reliability methods and their mutual differences: (1) the variation of the item variances, (2) the dimensionality due to the correlational structure, and (3) the strength of the inter-item correlations. To investigate the effects of these factors, computational studies were used.

We performed four computational studies addressing the effect on reliability methods of: (1) Size of equal item variances and equal inter-item correlations representing one-dimensional item structures. This setup was a benchmark for the next three studies; (2) Spread of item variances while keeping inter-item correlations equal representing one-dimensional item structures; (3) Varying correlations between items from two different dimensions while correlations between items within dimensions were fixed. Results were presented for both correlation and covariance matrices; and (4) Varying within-dimension inter-item correlations while between-dimension inter-item correlations were fixed.

This article is organized as follows. We briefly discuss CTT (Lord & Novick 1968). The λ coefficients have been studied by Jackson and Agunwamba (1977). We briefly reiterate their line of reasoning as it greatly helps to understand each of the reliability methods and their mutual relationships. Next, we discuss the research method for the computational studies followed by the results, and finish by discussing the results and their implications for follow-up research.

2 Classical Test Theory

According to CTT, observed test score X can be decomposed into a true score T and measurement error E, such that $X = T + E$. Suppose a test consists of J items. Let

X_j be the score on item j, hence $X = \Sigma_{j=1}^{J} X_j$. Let σ_Y^2 denote the variance of Y, then it follows from the assumptions of classical test theory that the test-score reliability is defined as

$$\rho = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}. \tag{1}$$

In this definition, both σ_T^2 and σ_E^2 are unobservable, and this was the reason why psychometricians proposed several methods to approximate ρ on the basis of the inter-item covariance matrix obtained in a single test administration. Because of its brevity, in this article we use the notation of Jackson and Agunwamba (1977) which we introduce first.

Let $\sigma_{jj} = \sigma_{X_j}^2$ denote the observable item-score variance, σ_{jk} the inter-item covariance between items j and k, $t_j = \sigma_{T_j}^2$ the item true-score variance, $\sigma_{T_j T_k}$ the inter-item true-score covariance, $\theta_j = \sigma_{E_j}^2$ the item measurement-error variance, and $\sigma_{E_j E_k}$ the inter-item measurement-error covariance. Notice that $\sigma_{T_j T_k} = \sigma_{jk}$ and $\sigma_{E_j E_k} = 0$, for all pairs $j \neq k$. Covariance matrices $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_T$ are $J \times J$ symmetrical matrices, whereas $\mathbf{\Sigma}_E$ is a $J \times J$ diagonal matrix. Matrix $\mathbf{\Sigma}_X$ is positive definite (pd), meaning that for any vector $\mathbf{u}$ of size J, we have $\mathbf{u}'\mathbf{\Sigma}_X\mathbf{u} > 0$ (i.e., the determinant of $\mathbf{\Sigma}_X$ is positive), and $\mathbf{\Sigma}_T$ and $\mathbf{\Sigma}_E$ are positive semi-definite (psd), that is, $\mathbf{u}'\mathbf{\Sigma}_T\mathbf{u} \geq 0$ and $\mathbf{u}'\mathbf{\Sigma}_E\mathbf{u} \geq 0$ (i.e., the two matrices' determinants are non-negative). It may be noted that

$$\mathbf{\Sigma}_X = \mathbf{\Sigma}_T + \mathbf{\Sigma}_E. \tag{2}$$

For example, for $J = 4$ Eq. (2) equals

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{pmatrix} = \begin{pmatrix} t_1 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & t_2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & t_3 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & t_4 \end{pmatrix} + \begin{pmatrix} \theta_1 & 0 & 0 & 0 \\ 0 & \theta_2 & 0 & 0 \\ 0 & 0 & \theta_3 & 0 \\ 0 & 0 & 0 & \theta_4 \end{pmatrix}. \tag{3}$$

Let $\mathbf{1}$ denote the unity vector of size J, then $\sigma_T^2 = \mathbf{1}'\mathbf{\Sigma}_T\mathbf{1}$, $\sigma_E^2 = \mathbf{1}'\mathbf{\Sigma}_E\mathbf{1}$, and the reliability definition in Eq. (1) may be written as

$$\rho = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\mathbf{1}'\mathbf{\Sigma}_T\mathbf{1}}{\sigma_X^2} = 1 - \frac{\mathbf{1}'\mathbf{\Sigma}_E\mathbf{1}}{\sigma_X^2} = 1 - \frac{\sum \theta_j}{\sigma_X^2}. \tag{4}$$

3 Guttman's Reliability Coefficients and the GLB

3.1 Guttman's Reliability Coefficients

The six λ coefficients (Guttman 1945) are lower bounds to the reliability. Each is derived from necessary (not: sufficient; hence, none provides the GLB) conditions for $\mathbf{\Sigma}_T$ to be psd. Jackson and Agunwamba (1977) provided derivations of these lower bounds, also illuminating some of the mutual relationships between the six λs and their relationship to the reliability. For each λ lower bound, we explain its formal basis and its definition, but we refer the reader to the original sources for more details about the steps leading from the formal basis to the specific λ definition. Logically, the definition of λ_3 precedes the discussion of λ_2.

Coefficient $\boldsymbol{\lambda}_1$. Because CTT also applies to item scores, we know that $\sigma_{jj} = t_j + \theta_j$, and because $t_j \geq 0$, it follows that $\theta_j \leq \sigma_{jj}$; thus, $\sum \theta_j \leq \sum \sigma_{jj}$. Hence, a most simple lower bound to the reliability is

$$\lambda_1 = 1 - \frac{\sum \sigma_{jj}}{\sigma_X^2}. \tag{5}$$

Coefficient λ_1 only exploits the information from $\mathbf{\Sigma}_T$ that $t_j \geq 0$ $(j = 1, \ldots, J)$. The other five λ coefficients extract more information from $\mathbf{\Sigma}_T$.

Coefficient $\boldsymbol{\lambda}_3$. Coefficient λ_3, also known as coefficient α (Cronbach 1951), uses information based on all 2×2 principal submatrices of $\mathbf{\Sigma}_T$, with diagonal elements t_j and t_k, and off-diagonal elements σ_{jk} and σ_{kj}; in particular, coefficient λ_3 uses the property

$$0 \leq \sigma^2_{X_j - X_k} = t_j + t_k - 2\sigma_{jk}, j \neq k. \tag{6}$$

Combining the sums for all $j \neq k$, λ_3 can be derived to be equal to

$$\lambda_3 = \lambda_1 + J^{-1}\lambda_1, \tag{7}$$

which is readily rewritten in its well known form as

$$\lambda_3 = \alpha = \frac{J}{J-1}\left(1 - \frac{\sum \sigma_{jj}}{\sigma_X^2}\right). \tag{8}$$

Coefficient λ_3 is a lower bound for reliability, ρ. Eq. (7) shows that as $J \to \infty$, we find that $\lambda_3 \to \lambda_1$. Obviously, for small J, λ_3 will clearly exceed λ_1 but the difference soon becomes smaller. However, for a 20-item test for which $\lambda_1 = 0.80$, we find that $\lambda_3 = 0.80 + 0.80/20 = 0.84$, a difference that still is worthwhile to report.

Coefficient $\boldsymbol{\lambda}_2$. Each of the 2×2 principal submatrices of $\mathbf{\Sigma}_T$ needs to be psd, but coefficient λ_3 does not use this property. The psd property implies that the submatrices' determinants must be non-negative; that is

$$t_j t_k \geq \sigma_{jk}^2, j \neq k, \tag{9}$$

and this result is used to derive

$$\lambda_2 = \lambda_1 + \frac{\left(J(J-1)^{-1} \sum \sum_{j \neq k} \sigma_{jk}^2\right)^{1/2}}{\sigma_X^2}. \tag{10}$$

Coefficient λ_2 is a lower bound for reliability, ρ, but usually it is not the greatest lower bound.

Coefficient $\boldsymbol{\lambda}_4$. Coefficient λ_4 exploits the information in $\boldsymbol{\Sigma}_T$ as follows. Let $\mathbf{u}$ be a J-vector with elements equal to either $+1$ or -1, and use the inequality $\mathbf{u}'\boldsymbol{\Sigma}_T\mathbf{u} \geq 0$. The effect of vector $\mathbf{u}$ in the quadratic form is to produce a sum of the J variances t_j and $J(J-1)$ covariances σ_{jk} in which each term has either positive or negative signs. For example, vector $\mathbf{u}' = (-1 \ +1 \ -1)'$ produces $\mathbf{u}'\boldsymbol{\Sigma}_T\mathbf{u} = t_1 + t_2 + t_3 - 2\sigma_{12} + 2\sigma_{13} - 2\sigma_{23} \geq 0$, and replacing $\mathbf{u}$ by $-\mathbf{u}$ yields the same result. Vector $\mathbf{u} = (+1 + 1 + 1)'$ yields $\mathbf{u}'\boldsymbol{\Sigma}_T\mathbf{u} = \sum_j t_j + \sum \sum_{j \neq k} \sigma_{jk} = \sigma_T^2 \geq 0$. Because each element of $\mathbf{u}$ has one of two possible values, in total 2^J different vectors $\mathbf{u}$ are possible, but because the effect of $\mathbf{u}$ and $-\mathbf{u}$ on the matrix product is the same, one may consider only 2^{J-1} vectors.

Starting from Eq. (2), one may write for any vector $\mathbf{u}$,

$$\mathbf{u}'\boldsymbol{\Sigma}_X\mathbf{u} = \mathbf{u}'\boldsymbol{\Sigma}_T\mathbf{u} + \mathbf{u}'\boldsymbol{\Sigma}_E\mathbf{u}. \tag{11}$$

Because $\mathbf{u}'\boldsymbol{\Sigma}_T\mathbf{u} \geq 0$, it follows that

$$\mathbf{u}'\boldsymbol{\Sigma}_E\mathbf{u} = \sum \theta_j \leq \mathbf{u}'\boldsymbol{\Sigma}_X\mathbf{u}. \tag{12}$$

Substituting $\mathbf{u}'\boldsymbol{\Sigma}_X\mathbf{u}$ for $\sum \theta_j$ in Eq. (4), a lower bound for the reliability ρ is obtained by finding $\mathbf{u}$ such that

$$\lambda_4 = \max_{\mathbf{u}} \left(1 - \frac{\mathbf{u}'\boldsymbol{\Sigma}_X\mathbf{u}}{\sigma_X^2}\right). \tag{13}$$

Jackson and Agunwamba (1977) showed that coefficient λ_4 equals the maximum value of coefficient λ_3 (equal to α) for the split of the test into two subtests that have test scores Y_1 and Y_2 (i.e., two test parts, not necessarily of equal length, such that $X = Y_1 + Y_2, J = 2$), based on the items that correspond to the positive and the negative elements in $\mathbf{u}$, respectively.

Coefficient $\boldsymbol{\lambda}_5$. Coefficient λ_5 exploits the result that covariances which are relatively large in absolute value put the main constraint on $\sum t_j$. Arbitrarily, let us assume that column k of $\boldsymbol{\Sigma}_X$ contains the relatively large covariances. Then, based on $t_j t_k \geq \sigma_{jk}^2$ [Eq. (9)] one can deduce

$$\sum t_j \geq 2\left(\sum_{j\neq k}\sigma_{jk}^2\right)^{1/2}. \tag{14}$$

Because CTT holds for individual items, so that $X_j = T_j + E_j$, and hence for the item-variance decomposition [Eq. (2)], such that $\sigma_{jj} = t_j + \theta_j$, it is also true that

$$\sum \sigma_{jj} = \sum t_j + \sum \theta_j. \tag{15}$$

Substituting $\sum t_j$ in Eq. (14) by the right-hand side of Eq. (15), and rearranging the terms produces

$$\sum \theta_j \leq \sum \sigma_{jj} - 2\left(\sum_{j\neq k}\sigma_{jk}^2\right)^{1/2}. \tag{16}$$

Then, substituting $\sum \theta_j$ in the numerator of Eq. (4) by the right-hand side of Eq. (16) yields a lower bound for ρ. To find the greatest value for λ_5, one determines Eq. (16) for each of the J columns in $\mathbf{\Sigma}_T$, thus letting k play the role of index (i.e., $k = 1, \ldots, J$), and defines coefficient λ_5 as

$$\lambda_5 = \lambda_1 + \max_k \frac{2\left(\sum_{j\neq k}\sigma_{jk}^2\right)^{1/2}}{\sigma_X^2}. \tag{17}$$

Coefficient $\boldsymbol{\lambda}_6$. Coefficient λ_6 is based on the multiple regression of an item score X_j on the other $J-1$ item scores. Let matrix $\mathbf{\Sigma}_{jj}$ denote the $(J-1)\times(J-1)$ covariance matrix without row j and column j, and let $\boldsymbol{\sigma}_j' = (\sigma_{jk})'$, $k \neq j$, denote the $J-1$ vector containing covariances involving item j but not σ_{jj}. Then, it can be shown that the residual variance $\epsilon_j^2 = \sigma_{jj} - \boldsymbol{\sigma}_j'\mathbf{\Sigma}_{jj}^{-1}\boldsymbol{\sigma}_j$, and that this provides an upper bound for measurement error, θ_j; that is, $\theta_j \leq \epsilon_j^2$. Substituting θ_j in the numerator of Eq. (4) by ϵ_j^2 yields lower bound

$$\lambda_6 = 1 - \frac{\sum \epsilon_j^2}{\sigma_X^2}. \tag{18}$$

3.1.0.1 The Greatest Lower Bound The observed covariance matrix $\mathbf{\Sigma}_X$ can be produced by many different matrices $\mathbf{\Sigma}_T$ and $\mathbf{\Sigma}_E$. The GLB (Bentler & Woodward 1980) is computed using an algorithm that maximizes the estimate of the trace of measurement-error matrix $\mathbf{\Sigma}_E$. The resulting matrix $\widetilde{\mathbf{\Sigma}}_E$ and its complement $\widetilde{\mathbf{\Sigma}}_T$, must be positive semi-definite such that $\mathbf{\Sigma}_X = \widetilde{\mathbf{\Sigma}}_T + \widetilde{\mathbf{\Sigma}}_E$. The GLB estimates reliability formulated as $\rho = 1 - \frac{\sigma_E^2}{\sigma_X^2}$ [Eq. (1)]. Using the assumptions of CTT, reliability can be written as

$$\rho = 1 - \frac{tr(\boldsymbol{\Sigma}_E)}{\sigma_X^2}. \tag{19}$$

The GLB is obtained by replacing $tr(\boldsymbol{\Sigma}_E)$ with $tr(\widetilde{\boldsymbol{\Sigma}}_E)$, resulting in

$$GLB = 1 - \frac{tr(\widetilde{\boldsymbol{\Sigma}}_E)}{\sigma_X^2}. \tag{20}$$

If all items in the test are essentially tau-equivalent (Lord & Novick 1968p. 90), the GLB is equal to the reliability; that is, GLB $= \rho$. The GLB provides the worst-case scenario for the reliability given the covariance matrice $\boldsymbol{\Sigma}_X$ (Sijtsma 2009). There are multiple ways to estimate the GLB. For details we refer to Bentler and Woodward (1980) and Ten Berge et al. (1981). We used the function glb.algebraic from the *psych* r-package (Revelle 2015) to obtain the GLB.

3.2 Relations Between Methods

We reiterate the relationships between the six λ coefficients and the GLB, and between these methods and reliability ρ.

1. $\lambda_1 \leq \lambda_3 \leq \lambda_2$ (proof, see Jackson & Agunwamba 1977). For finite J, $\lambda_1 < \lambda_3$; for $J \to \infty$, $\lambda_1 = \lambda_3$, but this does not happen in practice. Furthermore, $\lambda_3 \leq \lambda_2 \leq GLB \leq \rho$ with equality if the items are essentially tau-equivalent (Lord & Novick 1968p. 50); that is, $T = T + a_{jk}$, where a_{jk} is a scalar. In this case, coefficient λ_1 falls short of ρ by a factor $(J-1)/J$, because $\lambda_1 = [(J-1)/J]\lambda_3$.
2. Lord and Novick (1968pp. 93–94) showed that coefficient λ_4 is a higher lower bound than coefficient λ_3; that is, $\lambda_3 = \alpha \leq \lambda_4$. Jackson and Agunwamba (1977) derived the conditions for which $\lambda_4 = GLB$. In particular, if $\boldsymbol{\nu} = (\nu_1, \ldots \nu_J)$ is the vector with elements $\nu = -1, +1$ that maximizes λ_4, then the authors prove that: If $\lambda_4 = GLB$, then it must hold that $\theta_j = \nu_j(\boldsymbol{\Sigma}_X\boldsymbol{\nu})_j$, all j, where $(\boldsymbol{\Sigma}_X\boldsymbol{\nu})_j$ denotes the jth element of the column vector $\boldsymbol{\Sigma}_X\boldsymbol{\nu}$, provided (1) $\boldsymbol{\Sigma}_T = \boldsymbol{\Sigma}_X - \sum \theta_j$ is psd, and (2) $\theta_j \geq 0$, all j. If these conditions are satisfied, $\lambda_4 = GLB$.

   For real-data problems one has to check whether the GLB solution provides error variances $\theta_j = \nu_j(\boldsymbol{\Sigma}_X\boldsymbol{\nu})_j$, all j; the latter quantities $\nu_j(\boldsymbol{\Sigma}_X\boldsymbol{\nu})_j$ can be derived from $\boldsymbol{\nu}$, the item weights vector that produced λ_4, and the observable covariance matrix $\boldsymbol{\Sigma}_X$. The authors noticed that in general coefficient λ_4 is a lower bound for ρ but that it does not equal the GLB.
3. Jackson and Agunwamba (1977) derived conditions for which $\lambda_2 < \lambda_5$ and $\lambda_4 < \lambda_5$, but these inequalities are not true in general.

We are unaware of other relationships that have been demonstrated between the six λ coefficients. In addition to algebraic relations, considering graphically displayed relations based on a computational study may be worthwhile, because (1) the mutual

relations between the λs are unknown for several λ pairs or only known under particular conditions that often are unfulfilled, and (2) except for λ_1 and λ_3, it is unknown how far the different λ values are apart, how far they are apart from the GLB, and whether differences are large enough to be of practical interest. In the next section, we discuss and present results of such a computational study.

4 Method

In this section, we discuss the setup of four computational studies. The purpose of these studies was to explore how different factors influence the distance between the coefficients λ_1 through λ_6 in relation to the GLB, and to study the numerical differences among λ coefficients. In each computational study, λ_1 through λ_6 and the GLB were computed using correlation matrices and covariance matrices. The λ coefficients provided values at most equal to the GLB.

Study 1 was a benchmark for the other studies. In Study 1, the values of the seven methods were computed using constructed correlation matrices containing equal item variances and equal inter-item correlations, which may follow from essential tau-equivalence. In Study 2, the item-score variances were varied while keeping the inter-item correlations equal and fixed. In studies 3 and 4, the item-score variances were varied, and inter-item correlations were varied so as to construct a two-dimensional item structure. Thus, the effect of multi-dimensionality on the reliability methods could be studied. In Study 3, correlations between items from different dimensions were varied and correlations between items within dimensions were fixed. In Study 4, by contrast, between-dimension inter-item correlations were fixed and within-dimension inter-item correlations were varied.

4.1 Study 1: Equal Correlations

For 4, 6, and 8 standardized items, we investigated inter-item correlation matrices in which all inter-item correlations ρ_{jk} were equal. Different correlation matrices were constructed by letting ρ_{jk} run from 0 to 1 in steps equal to 0.025; that is, $\rho_{jk} = 0, 0.025, \ldots, 1$. This resulted in 41 correlation matrices $\mathbf{R}_n$ ($n = 1, \ldots, 41$). For each of three test lengths, $J = 4, 6, 8$, λs were computed for each of the 41 correlation matrices. For example, for $J = 4$ the correlation matrices equaled

$$\mathbf{R}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{R}_2 = \begin{pmatrix} 1 & 0.025 & 0.025 & 0.025 \\ 0.025 & 1 & 0.025 & 0.025 \\ 0.025 & 0.025 & 1 & 0.025 \\ 0.025 & 0.025 & 0.025 & 1 \end{pmatrix},$$

$$\ldots, \quad \mathbf{R}_{41} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

All items are standardized (i.e., $\mu_j = 0$ and $\sigma_{jj} = 1$), so that $\mathbf{\Sigma}_{Xn} = \mathbf{R}_n$.

4.2 Study 2: Varying Item-Score Variances

To isolate the effect of variation of item-score variances, inter-item correlations were fixed at $\rho_{jk} = 0.3$, which is a value typical of empirical test research. For example, the NEO-PR big-five personality inventory (McCrae & Costa 1999, retrieved from *Psych* package, Revelle 2015) reports mean inter-item correlations within each of the five facets equal to approximately 0.3. Test length equaled $J = 4, 6, 8$. For example, for $J = 4$, the correlation matrix $\mathbf{R}$ equaled

$$\mathbf{R} = \begin{pmatrix} 1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1 \end{pmatrix}. \tag{21}$$

Covariance matrices were constructed as follows. For $j = 1, 2, 3$, we chose $\sigma_{jj} = 1.5$, which is representative of 5-point Likert scales regularly found in psychological research. Across different covariance matrices, σ_{44} varied from 0.25 to 4.00 in steps equal to 0.05; that is, $\sigma_{44} = 0.25, 0.30, \ldots, 4.00$, resulting in 76 covariance matrices. Covariance matrices equaled

$$\mathbf{\Sigma}_{X_1} \approx \begin{pmatrix} 1.5 & 0.45 & 0.45 & 0.18 \\ 0.45 & 1.5 & 0.45 & 0.18 \\ 0.45 & 0.45 & 1.5 & 0.18 \\ 0.18 & 0.18 & 0.18 & 0.25 \end{pmatrix}, \quad \mathbf{\Sigma}_{X_2} \approx \begin{pmatrix} 1.5 & 0.45 & 0.45 & 0.20 \\ 0.45 & 1.5 & 0.45 & 0.20 \\ 0.45 & 0.45 & 1.5 & 0.20 \\ 0.20 & 0.20 & 0.20 & 0.30 \end{pmatrix},$$

$$\ldots, \quad \mathbf{\Sigma}_{X_{76}} \approx \begin{pmatrix} 1.5 & 0.45 & 0.45 & 0.73 \\ 0.45 & 1.5 & 0.45 & 0.73 \\ 0.45 & 0.45 & 1.5 & 0.73 \\ 0.73 & 0.73 & 0.73 & 4 \end{pmatrix}.$$

For 6 and 8 items, inter-item correlations equaled those used in [Eq. (21)]; that is, $\rho_{jk} = 0.3$. For $J = 6$, for the first four items, item-score variance $\sigma_{jj} = 1.5$ ($j = 1, \ldots, 4$), and for the last two items 5 and 6, σ_{55} and σ_{66} varied by increasing steps equal to 0.05, so that $\sigma_{55} = \sigma_{66} = c$, with $c = 0.25, 0.30, \ldots, 4.00$. For $J = 8$, the same numerical choices were made, keeping σ_{11} through σ_{55} equal to 1.5, and varying σ_{66}, σ_{77}, and σ_{88} by increasing steps equal to 0.05, starting at 0.25 and ending with 4.00, so that $\sigma_{66} = \sigma_{77} = \sigma_{88} = c$, with $c = 0.25, 0.30, \ldots, 4.00$.

4.3 Study 3: Two Dimensions, Varying Correlations Between Dimensions

In this example, the correlation matrices had a two-dimensional structure. Inter-item correlations were manipulated such that for different matrices the two dimensions either were negatively related, unrelated, positively related, or indistinguishable, thus representing one dimension. The range of correlations between items from different dimensions was based on the range of correlations between items from different facets of the NEO-PI-R available in the *Psych* package (Revelle 2015). In the first condition item variances were equal to 1, so the correlation and covariance matrix were equal. In the second condition the item variance varied. In the second example covariance matrices were based on the correlation matrices in the first example. This was done so as to create matrices resembling those found in empirical research.

For 4, 6, and 8 items, the inter-item correlations and the item variances were manipulated, resulting in 37 correlation and covariance matrices. Consider $J = 4$: A two-dimensional structure was constructed by dividing correlation matrix $\mathbf{R}$ into block $\mathbf{A}$ and block $\mathbf{B}$ [Eq. (22)], such that

$$\mathbf{R} = \left(\begin{array}{cc|cc} \rho_{11} & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{21} & \rho_{22} & \rho_{23} & \rho_{24} \\ \hline \rho_{31} & \rho_{32} & \rho_{33} & \rho_{34} \\ \rho_{41} & \rho_{42} & \rho_{43} & \rho_{44} \end{array}\right) = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{B} & \mathbf{A} \end{array}\right). \tag{22}$$

For all matrices, block $\mathbf{A}$ was a $J/2 \times J/2$ matrix with diagonal elements equal to 1 and off-diagonal elements equal to 0.6. Block $\mathbf{B}$ differed for all matrices. Matrix $\mathbf{B}_n$ ($n = 1, \ldots, 37$) is a $J/2 \times J/2$ matrix with all off-diagonal elements equal to $\rho_{jk(n)} = -0.3225 + 0.0225 \times n$. For example, for $J = 4$ the correlation matrices $\mathbf{R}_n$ equaled

$$\mathbf{R}_1 = \begin{pmatrix} 1 & 0.6 & -0.3 & -0.3 \\ 0.6 & 1 & -0.3 & -0.3 \\ -0.3 & -0.3 & 1 & 0.6 \\ -0.3 & -0.3 & 0.6 & 1 \end{pmatrix}, \quad \mathbf{R}_2 \approx \begin{pmatrix} 1 & 0.6 & -0.28 & -0.28 \\ 0.6 & 1 & -0.28 & -0.28 \\ -0.28 & -0.28 & 1 & 0.6 \\ -0.28 & -0.28 & 0.6 & 1 \end{pmatrix},$$

$$\ldots, \quad \mathbf{R}_{37} = \begin{pmatrix} 1 & 0.6 & 0.6 & 0.6 \\ 0.6 & 1 & 0.6 & 0.6 \\ 0.6 & 0.6 & 1 & 0.6 \\ 0.6 & 0.6 & 0.6 & 1 \end{pmatrix}.$$

Correlation matrices for 6 and 8 items were constructed by extending blocks $\mathbf{A}$ and $\mathbf{B}$ by 1 or 2 rows and columns, respectively.

Covariance matrices were constructed as follows. For the three test lengths, the variances of the even numbered items equaled 1 and the variances of the odd

numbered items equaled 2. For $J = 4$, we constructed 37 covariance matrices that equaled

$$\boldsymbol{\Sigma}_{X_1} \approx \begin{pmatrix} 2 & 0.85 & -0.6 & -0.6 \\ 0.85 & 1 & -0.6 & -0.6 \\ -0.6 & -0.6 & 2 & 0.85 \\ -0.6 & -0.6 & 0.85 & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_{X_2} \approx \begin{pmatrix} 2 & 0.85 & -0.56 & -0.56 \\ 0.85 & 1 & -0.56 & -0.56 \\ -0.56 & 0.56 & 2 & 0.85 \\ -0.56 & 0.56 & 0.85 & 1 \end{pmatrix},$$

$$\ldots, \quad \boldsymbol{\Sigma}_{X_{37}} \approx \begin{pmatrix} 2 & 0.85 & 1.2 & 1.2 \\ 0.85 & 1 & 1.2 & 1.2 \\ 1.2 & 1.2 & 2 & 0.85 \\ 1.2 & 1.2 & 0.85 & 1 \end{pmatrix}.$$

4.4 Study 4: Two Dimensions, Varying Correlations Within Dimensions

As in Study 3, the coefficients were calculated using covariance matrices belonging to one of two conditions. In the first condition item variances were all equal to 1, so the correlation matrix and covariance matrix were equal. In the second condition the item variance varied. Again, the correlation matrices were divided into blocks A and B [Eq. (22)] but the inter-item correlations between the dimensions (block B in Study 3) were fixed and the inter-item correlations within the dimensions (block A in Study 3) were varied. This resulted in a $J/2 \times J/2$ matrix $\mathbf{A}_n (n = 1, \ldots, 41)$, with diagonal elements 1 and off-diagonal elements $\rho_n = -0.025 + 0.025 \times n$. All the elements of the $J/2 \times J/2$ matrix $\mathbf{B}$ are 0.1. This resulted in correlation matrices

$$\mathbf{R}_n = \left(\begin{array}{c|c} \mathbf{A}_n & \mathbf{B} \\ \hline \mathbf{B} & \mathbf{A}_n. \end{array}\right), \quad \text{for} \quad n = 1, \ldots, 41. \tag{23}$$

For $J = 4$, the correlation matrices $\mathbf{R}_n$ equaled

$$\mathbf{R}_1 = \begin{pmatrix} 1 & 0 & 0.1 & 0.1 \\ 0 & 1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 1 & 0 \\ 0.1 & 0.1 & 0 & 1 \end{pmatrix}, \quad \mathbf{R}_2 = \begin{pmatrix} 1 & 0.025 & 0.1 & 0.1 \\ 0.025 & 1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 1 & 0.025 \\ 0.1 & 0.1 & 0.025 & 1 \end{pmatrix},$$

$$\ldots, \quad \mathbf{R}_{41} = \begin{pmatrix} 1 & 1 & 0.1 & 0.1 \\ 1 & 1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 1 & 1 \\ 0.1 & 0.1 & 1 & 1 \end{pmatrix}.$$

Correlation matrices for $J = 6, 8$ were obtained by adding one or two rows and columns to blocks **A** and **B**, respectively.

Covariance matrices were constructed using the correlation matrices. Similar to Study 3, for the even numbered items the item-score variances equaled 1 and for the odd numbered items the item-score variances equaled 2. Using the 41 correlation matrices and the item-score variances, 41 covariance matrices $\mathbf{\Sigma}_{X_n}$ were constructed. For example, for $J = 4$, the matrices equaled

$$\mathbf{\Sigma}_{X_1} \approx \begin{pmatrix} 2 & 0 & 0.2 & 0.14 \\ 0 & 1 & 0.14 & 0.1 \\ 0.2 & 0.14 & 2 & 0 \\ 0.14 & 0.1 & 0 & 1 \end{pmatrix}, \quad \mathbf{\Sigma}_{X_2} \approx \begin{pmatrix} 2 & 0.04 & 0.2 & 0.14 \\ 0.04 & 1 & 0.14 & 0.1 \\ 0.2 & 0.14 & 2 & 0.04 \\ 0.14 & 0.1 & 0.04 & 1 \end{pmatrix},$$

$$\ldots, \quad \mathbf{\Sigma}_{X_{41}} \approx \begin{pmatrix} 2 & 1.41 & 0.2 & 0.14 \\ 1.41 & 1 & 0.14 & 0.1 \\ 0.2 & 0.14 & 2 & 1.41 \\ 0.14 & 0.1 & 1.41 & 1 \end{pmatrix}.$$

5 Results

5.1 Study 1: Equal Correlations

Figure 1 (left panel) shows that for fixed test length, reliability increases as inter-item correlations, ρ_{jk}, increase. This increase is faster for longer tests. By definition, λ_1 produced the lowest values of the λs, and the GLB produced the highest value. Because in each matrix **R** all inter-item correlations were equal, a necessary condition for essential tau-equivalence was satisfied; hence, λ_2, λ_3, λ_4 and GLB provided the same values. Equal inter-item correlations do not imply essential tau-equivalence; hence, λ_2, λ_3, λ_4 and GLB do not necessarily provide the reliability, ρ. At best λ_5 and λ_6 produced values that were lower than GLB by 0.04 and 0.01 units, respectively.

The difference between λ_6 on the one hand and λ_2, λ_3, λ_4 and the GLB on the other hand was smallest for the lowest and highest values of ρ_{jk}. As inter-item correlation ρ_{jk} increased, the difference between λ_1 and λ_5 on the one hand, and the GLB on the other hand increased. Method λ_5 was only closer to the GLB than λ_6 (at most by 0.01 units) for lower values of ρ_{jk} and the difference was greater as fewer items were used. When $\rho_{jk} = 1$, matrix **R** had determinant equal to 0; hence, λ_6 which uses the multiple regression model could not be computed.

For this study and the next three studies, method λ_1 not only was furthest from the GLB, but the distance was so large that λ_1 was useless compared to the other λs. Therefore, there is no discussion of the results for λ_1 in the remainder of this section. Results for λ_1 can be found in all figures.

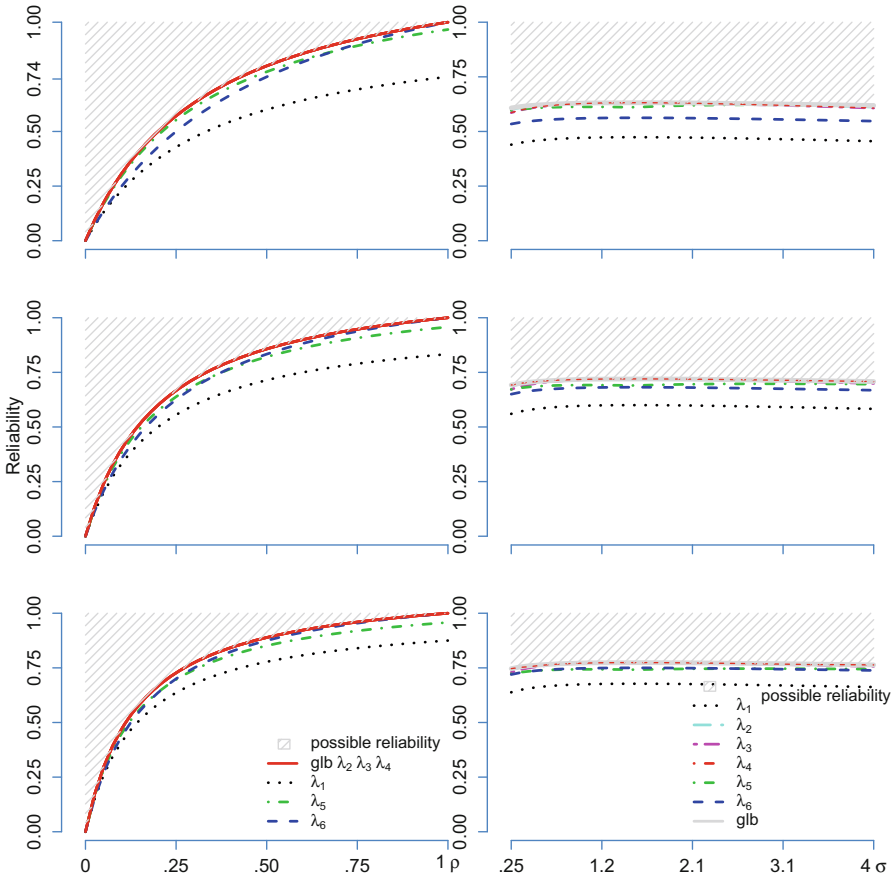


Fig. 1 Reliability coefficients as function of inter-item correlation ρ, or item variance σ, for $J = 4$ (*top*), $J = 6$ (*middle*), and $J = 8$ (*bottom*), with equal correlations (*left*) and varying item variances (*right*)

5.2 *Study 2: Varying Item-Score Variances*

Figure 1 (right panel) shows that the effect of manipulating the item variances on the differences between the λs and the GLB was small. The differences were approximately equal to the differences found in Study 1 for $\rho_{jk} = 0.3$. λ_2, λ_3, and λ_4 almost always yielded higher values than λ_5 and λ_6, except for a few conditions discussed in the next paragraph. λ_2, λ_3, and λ_4 differed equally from the GLB, but the difference was negligible, and was always smaller than 0.02. For $J = 4$, when the item variances differed the most, λ_2 produced slightly higher values than the other methods.

For $J = 4$, the four covariance matrices having the most extreme item-score variance (i.e., $\sigma_{44} = 0.25, 0.30, 3.95, 4.00$) produced the smallest difference

between λ_5 and the GLB. The difference between λ_5 and the GLB was largest when item variances were equal. This results from λ_5 utilizing differences between columns of the covariance matrix to find the best possible estimate for item true-score variance (Verhelst 2000p. 7). Because the inter-item correlations in this study were equal, the differences between columns were smallest when item variances were identical.

Because the differences between methods λ_2 through λ_5 and the GLB were small, the effect of increasing test length was not clear-cut. For method λ_6, compared to manipulating item variance, increasing test length had a stronger effect. This can be understood from the regression model containing more predictors as tests grow longer, hence producing smaller residual item variances.

5.3 Study 3: Two Dimensions, Varying Correlations Between Dimensions

Figure 2 shows that for all λs the distance to the GLB was smaller as the inter-item correlations were more similar, thus causing the two-dimensional structure of the matrices to disappear. In most conditions, λ_4 was closest to the GLB (difference always < 0.08). Only when $J = 6$, all item variances equaled $\sigma_{jj} = 1$, and the between-dimension inter-item correlations were approximately $\rho_{jk} = 0$, the difference between λ_6 and the GLB was smaller than the difference between λ_4 and the GLB (at most 0.01).

In most conditions, λ_3 differed the most from the GLB. When all inter-item correlations were equal (i.e., $\rho_{jk} = 0.6$), it holds that $\lambda_2 = \lambda_3 = \lambda_4 = GLB$. When ρ_{jk} approached 0.6 from below, λ_3 eventually was closer to the GLB than λ_5 and λ_6 (at most 0.03 and 0.04, respectively). Figure 2 shows that as test length increased, the λ_3 curve intersected with the λ_5 and λ_6 curves at lower ρ_{jk} values.

Coefficients λ_2, λ_5, and λ_6 all had similar distances to the GLB, with distances between λ coefficients being more extreme as test length grew (Fig. 2). λ_6 was almost always closest to the GLB, except when $J = 4$ and approximately $\rho_{jk} = 0.6$. For all conditions, we found $\lambda_2 > \lambda_5$. Creating covariance matrices from the correlation matrices by increasing the variance of even numbered items by 1 was not sufficient to create a column in the covariance matrix with a sum of squared covariances larger than $\frac{J^2}{4}$ times the mean item variance (Verhelst 2000p. 8).

Differences between results from correlation matrices and results from covariance matrices were small. The two most noticeable differences were found for $J = 6$. The difference between both λ_4 and λ_5 and the GLB were notably smaller (0.04 and 0.03, respectively). Increasing item variances by 1 for uneven items did not produce differences between the columns of the covariance matrices that were large enough to result in favorable results for λ_5.

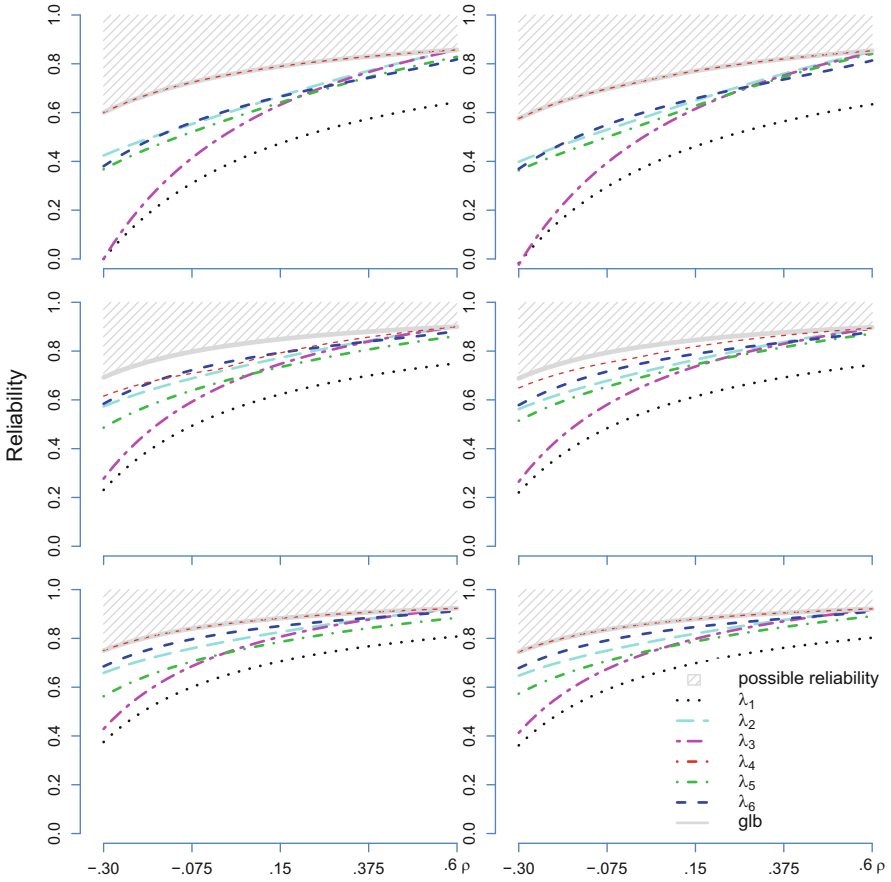


Fig. 2 Reliability coefficients for two-dimensional structure as a function of inter-item correlations (ρ) between dimensions, for $J = 4$ (*top*), $J = 6$ (*middle*), and $J = 8$ (*bottom*), with standardized items (*left*) and unstandardized items (*right*)

5.4 Study 4: Two Dimensions, Varying Correlations Within Dimensions

Figure 3 shows the results for the two-dimensional item structure when dimensions were weakly related. Similar to the previous studies, for most conditions λ_4 was closest to the GLB. Except when $J = 6$, for the top half of the within-dimension inter-item correlations (for inter-item correlations approximately larger 0.48), λ_6 outperformed λ_4. Compared to λ_4, λ_6 was closer to the GLB, and the difference between the λs and the GLB was greater as the correlation between dimensions increased (being 0.04 at its maximum). Also similar to Study 3, except for λ_5 differences between results for correlation matrices and covariance matrices were

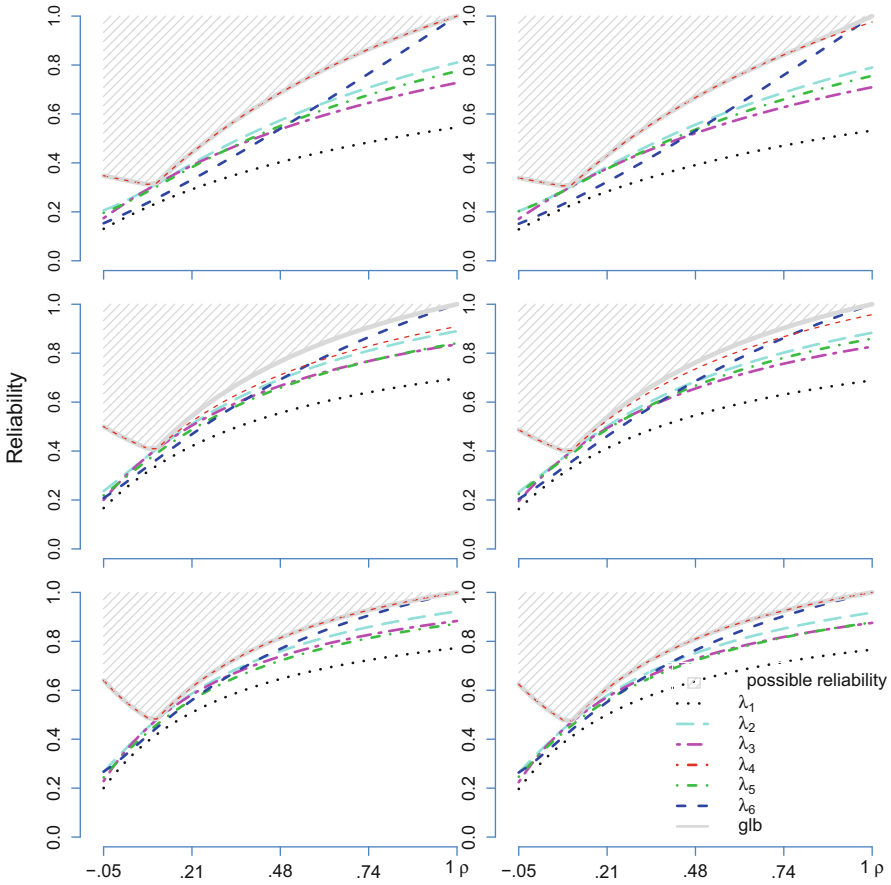


Fig. 3 Reliability coefficients for two-dimensional structure as a function of inter-item correlations (ρ) within dimensions, for $J = 4$ (*top*), $J = 6$ (*middle*), and $J = 8$ (*bottom*), with standardized items (*left*) and unstandardized items (*right*)

small. For $J = 6$ and $J = 8$, λ_5 produced higher values for the covariance matrices than for the correlation matrices but these higher values were not closer to the GLB than for example λ_4 and λ_6.

Of the remaining λs, λ_6 benefited most from higher within-dimension inter-item correlations. This result was found especially for the top half of the within-dimension inter-item correlations (again for inter-item correlations approximately larger than 0.48). Across all conditions, λ_2 was closer to the GLB than λ_3 and λ_5.

6 Discussion

None of the λs was closest to the GLB for all conditions discussed. However, compared to the other λs, in general method λ_4 was closest to the GLB. This result may have been facilitated by the structure of the correlation matrices that made selection of similar test halves easy. For 4 and 8 items and equal item variances this structure was perfect. Methods λ_1 and λ_3 are not serious competitors for the GLB. Method λ_1 not only is the smallest lower bound of the six λs but the difference with the other λs and the GLB is too large to be useful. Although generally much higher than λ_1, method λ_3 also appears rather useless, a result that has been discussed in different contexts (e.g., Cortina 1993; Cronbach 2004; Schmitt 1996; Sijtsma 2009; Zinbarg, Revelle, Yovel, & Li 2005).

Intuitively, method λ_5 might have been considered a good alternative to the GLB because of its capacity to cope with variation within the covariance matrix. However, even though the computational examples in this study may be considered rather representative of data structures typically encountered in psychological research, λ_5's performance was worse than that of the other methods (except λ_1). For all λs, in general differences between results for covariance matrices and correlation matrices caused by varying item variance were modest to small.

For small to moderate samples not containing more than 1000 cases, the GLB suffers from strong positive sampling bias (Ten Berge & Sočan 2004) and alternative methods may be considered. Candidates replacing the GLB for small to moderate samples are λ_2, λ_4 and λ_6. Only when differences in item variance are large and inter-item correlations are very similar is λ_5 a viable candidate. For λ_4 results are available showing bias is likely to be small for values greater than 0.85, test length smaller than 25 items and sample size greater than 3000 (Benton 2015). Research addressing the sampling variance of these methods is needed and we are currently studying this issue.

References

Bentler, P. M., & Woodward, J. A. (1980). Inequalities among lower bounds to reliability: With applications to test construction and factor analysis. *Psychometrika*, *45*, 249–267.

Benton, T. (2015). An empirical assessment of Guttman's lambda 4 reliability coefficient. In R. E. Millsap, D. M. Bolt, L. A. van der Ark, & W. -C. Wang (Eds.), *Quantitative psychology research: The 78th annual meeting of the Psychometric Society* (pp. 301–310). New York, NY: Springer.

Cortina, J. M. (1993). What is coefficient alpha? an examination of theory and applications. *Journal of Applied Psychology*, *78*, 98–104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*, 391–418.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*, 255–282.

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: Algebraic lower bounds. *Psychometrika*, *42*, 567–578.

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

McCrae, R. R., & Costa, P. T. (1999). A five-factor theory of personality. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 139–153). New York: Guilford Press.

Revelle, W. (2015). *Psych: Procedures for personality and psychological research Version 1.5.8 [computer software]*. Evanston, IL. Retrieved from http://CRAN.R-project.org/package=psych.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 350–353.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107–120.

Ten Berge, J. M. F., Snijders, T. A. B., & Zegers, F. E. (1981). Computational aspects of the greatest lower bound to the reliability and constrained minimum trace factor analysis. *Psychometrika*, *46*, 201–213.

Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, *69*, 613–625.

Verhelst, N. (2000). *Estimating the reliability of test from single test administration.* Unpublished report. Arnhem, The Netherlands: Cito. Retrieved from http://www.cito.com/research_and_development/psychometrics/~/media/cito_com/research_and_development/publications/cito_report98_2.ashx.

Woodhouse, B., & Jackson, P. H. (1977). Lower bounds for the reliability of a test composed of nonhomogeneous items II: A search procedure to locate the greatest lower bound. *Psychometrika*, *67*, 251–259.

Zinbarg, R., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_w: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*, 122–133.

Optimizing the Costs and GT based reliabilities of Large-scale Performance Assessments

Yon Soo Suh, Dasom Hwang, Meiling Quan, and Guemin Lee

Abstract In generalizability theory (GT), higher levels of reliability can be obtained by increasing facet sample sizes but at the expense of increasing expenditure and resources. The challenging task is identifying optimal sample sizes that balance such psychometric and practical considerations. As such, the objective of our research was to demonstrate the use of mixed integer nonlinear programming, an optimization procedure, in attaining the most cost-efficient measurement design subject to both psychometric and practical constraints. The optimization procedure was applied to the context of large-scale performance assessments where costs and reliability are important but conflicting issues. The results suggest that the optimization method can be a useful tool in determining the optimal sampling factors to achieving a desired reliability coefficient among multiple feasible solutions. Moreover, they demonstrate how practitioners not only face a trade-off between costs and desired reliability where costs increase exponentially in order to heighten reliability but also demonstrate the need for test developers to consider possible additional practical constraints along with budget and reliability such as restrictions on the number of students, tasks, raters or any other facet of interest.

Keywords Generalizability theory • Large-scale performance assessment • Mixed-integer nonlinear programming • Optimal sample sizes • Reliability

1 Introduction

Despite the many purposed advantages of performance assessments, technical quality and cost issues are often mentioned as obstacles to their adaptation to large scale settings (Darling-Hammond, Newton & Wei 2013). The former is related to issues of the reliability of performance assessments due to sampling variability or measurement error (Shavelson, Baxter & Gao 1993) and the latter involves increased costs because of higher task development, administration and

Y.S. Suh (✉) • D. Hwang • M. Quan • G. Lee
Department of Education, Yonsei University, Seoul, South Korea
e-mail: yssuh860909@gmail.com

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer
Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_13

rater costs following the complexity of the test format (Stecher & Klein 1997). Nonetheless, in an era of standards-based accountability and high-stakes testing, combined with technological developments and cost-saving measures, performance assessments are being re-examined (Darling-Hammond et al. 2013; Lane 2010). However, there is little literature on efficiently implementing such assessments while simultaneously considering issues of reliability, cost and other practical constraints. Also, there is little research targeted specifically towards school-level reliability, although it can differ from individual-level reliability to lead to misinterpretations (Gao, Shavelson & Baxter 1994; Jeon, Lee, Hwang & Kang 2009). As such, this study illustrates the integration of a cost optimization framework with generalizability theory (GT) to achieve the most cost-effective measurement design under pre-specified psychometric and practical constraints for large-scale performance assessments where school-level reliability is of concern.

2 Generalizability Theory

Generalizability theory (GT) provides a framework for identifying and estimating multiple possible sources of variability in a measurement when calculating reliability to accurately account for the underlying measurement structure of tests such as performance assessments. Furthermore, it can be applied to plan and decide future studies because GT allows researchers to implement different data collection designs and manipulate facet sample sizes to derive various alternative measurement designs and reliability estimates. GT consists of a two stage process with a distinction between generalizability (G) studies and decision (D) studies.

G-study A G-study addresses questions of how well measures taken in one context generalize to another by estimating the errors of measurement via decomposing an observed score into an overall mean and several effects and then obtaining their variance components. The target population is called the object of measurement and each set of characteristics that is a potential source of error is referred to as a facet of measurement. A universe of admissible observations is then defined by all possible combinations of conditions of the facets. The relative magnitudes of the estimated variance components associated with each facet and their interactions from the universe provide information about the potential sources of error.

D-study The variance components of a G-study are used to determine the generalizability of sampled observations to a universe of similar observations. In planning a D-study, the decision maker first defines the universe of generalization which contains those facets and conditions to generalize to and calculates the universe scores and its variance, universe-score variance, for the object of measurement as well as the appropriate error variances for the facets of interest. The ultimate purpose of a D-study is to provide summary coefficients analogous to the reliability coefficient in classical test theory. There are two kinds of coefficients: the generalizability coefficient for norm-referenced interpretations, the ratio of universe

score variance to itself and relative error variance ($E\rho^2 = \frac{\sigma^2(\tau)}{\sigma^2(\tau)+\sigma^2(\delta)}$), and the index of dependability for criterion-referenced interpretations, the ratio of universe score variance to itself and absolute error variance ($\Phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau)+\sigma^2(\Delta)}$). GT reliability coefficients can be manipulated by sampling along the facets to investigate the trajectory of change subject to different sample sizes so as to identify the optimal level of reliability in a D-study (Brennan 2001; Shavelson 1989).

3 Optimization Procedure

An optimal problem formulation creates a mathematical model of the optimization problem, which is solved using an optimization algorithm of choice. The outline of the steps usually involved in an optimization procedure is given in Fig. 1.

Step 1 involves identifying the underlying design variables important to the working of the optimization design while other design parameters remain fixed or vary in relation to them. Step 2 is finding the objective function which mathematically represents the purpose of optimization, in terms of a maximization or minimization function of the design variables and parameters. Step 3 is related to forming any possible constraints which represent functional relationships among the design variables and parameters that meet certain circumstances or resource limitations. Various constraints from single versus multiple; inequality versus equality; and linear versus nonlinear constraints exist. Step 4 is also an optional phase of constructing the lower and upper bounds of each design variable. The search algorithm locates the solutions within the feasible region surrounded by constraints as well as the bounds as these bounds are also a type of constraint. Step 5 and final task of the optimization procedure is running a search algorithm or calculation process which usually derives optimal solutions by way of an iterative process.

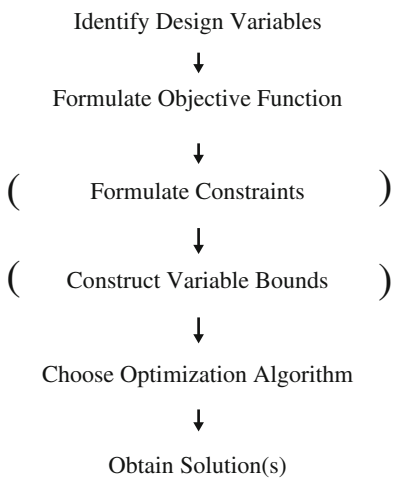


Fig. 1 Flowchart of optimization procedure

The mathematical formulation is

$$\begin{aligned} \boldsymbol{x} &= \{x_1, x_2, \ldots, x_n\} \\ \text{Minimize/Maximize } & f(\boldsymbol{x}) \\ \text{Subject to } & \boldsymbol{g}(\boldsymbol{x}) \\ \boldsymbol{x} \in R &= \left\{x_{i,lowerbounds} \leq x_i \leq x_{i,upperbounds} \ (i = 1, \ldots n)\right\} \end{aligned} \tag{1}$$

where $\boldsymbol{x}$ is a vector of design variables, $f(\boldsymbol{x})$ is the objective function, $\boldsymbol{g}(\boldsymbol{x})$ is a vector of constraints and R equals the feasible region (Antoniou & Lu 2007).

4 Optimization in Generalizability Theory

GT allows the flexibility of obtaining higher levels of generalizability by increasing facet sample sizes accordingly. However, facet sample sizes cannot be increased to infinity due to budget restrictions and other possible limits such as number of tasks and raters, which constricts the amount of measurement precision that is attainable. The obstacle in designing a measurement procedure is to pinpoint facet sample sizes that simultaneously produces acceptable reliability while keeping within the bounds of such constraints (Meyer, Liu & Mashburn 2013).

This problem is exacerbated in that GT considers multiple sources of measurement error as in the case of performance assessments so that various different combinations of the facet conditions can derive the same reliability, each at a different cost (Marcoulides & Goldstein 1990). Furthermore, the costs involved may not be proportional to the total number of observations in order to derive a higher reliability as in the case of the Spearmen-Brown prophecy formula for multiple-choice assessments (Marcoulides & Goldstein 1991). In other words, a smaller total number of observations can result in overall lower costs and higher reliability than a larger counterpart, which is counterintuitive.

The decision maker must balance all these considerations to choose the most appropriate D-study design. This can a tedious process involving a vast number of combinations to be prone to error and no guarantee of optimal results if done manually. Also, the D-study cannot directly take cost information into account which is problematic as costs cannot be automatically substituted with the number of observations (Parkes 2000). On the other hand, the incorporation of optimization techniques with GT makes it possible to achieve the most efficient allocation of resources to maximize reliability or minimize costs while accounting for such various concerns and thus procure both quality and economy of the measurement procedure in one analysis.

Two optimization procedures incorporating GT have been suggested so far: (1) maximize the generalizability coefficient (minimize relative error variance) under cost-constraints (Sanders, Theunissen & Baas 1991), or (2) minimize the cost

function under generalizability coefficient constraints (Peng, Li & Wan 2012). Such procedures have been adapted to single and multi-facet designs (Woodward & Joe 1973); crossed and nested designs; and univariate and multivariate designs (Marcoulides & Goldstein 1991); albeit scantly.

For both cases, variable bounds can be imposed as additional constraints, such that the sample size of a facet has to be at least some value or cannot exceed a certain number. These constraints are very likely in real testing situations which often have a set limit of testing time or a small number of possible raters because of the costs. It is to be noted that although such related extra constraints have been suggested (Marcoulides & Goldstein 1990) and despite their prevalence, there is very little research that has constructed specific variable bounds and compared the results from using these bounds with those without.

Another issue of optimization in GT is that most initial solutions derived from optimization algorithms are not integer values which are not possible for facet sample sizes. While Marcoulides and Goldstein (1991) recommended finding all possible permutations rounded to the nearest integer, this is labor intensive and can distort the results so that they are no longer optimal. As such, Sanders (1992) demonstrated the application the branch and bound algorithm to find optimal integer solutions.

5 Example Optimization Study

Data The proposed optimization procedure is demonstrated using data collected by Gao et al. (1994) from the California Assessment Program's (CAP) sixth-grade field test data on science performance assessment collected in 1990. The sample size was 600 students from 15 randomly selected students from 40 randomly selected schools. The assessment consisted of the five content domains of Electricity, Leaves, Rocks, Measurement, Acids and Bases. As a hands-on performance assessment, the students rotated between five self-contained stations where they performed relevant tasks at timed intervals. The scoring of each task was based on a holistic scoring rubric ranging from 0 (No attempt) to 4 (outstanding). Three raters with a rater agreement of at least 85 % rated the students.

G-study For this study, the primary object of measurement was schools (s). Based on the design of Gao et al. (1994), raters (r), tasks (t) and students (p) were the facets of variation. As such, the G study design was (person: school) × task × rater.

D-study D-study design of choice was also (Person:school) × Task × Rater as it was the most adequate representation of the assessment's measurement structure. The relative error ($\sigma^2(\delta)$) with schools as the objective of measurement and the corresponding generalizability coefficient were of interest. The relative error was

$$\sigma^2(\delta) = \frac{\sigma^2_{p:s}}{n'_{p:s}} + \frac{\sigma^2_{sr}}{n'_r} + \frac{\sigma^2_{st}}{n'_t} + \frac{\sigma^2_{(p:s)r}}{n'_{p:s}n'_r} + \frac{\sigma^2_{(p:s)t}}{n'_{p:s}n'_t} + \frac{\sigma^2_{srt}}{n'_r n'_t} + \frac{\sigma^2_{(p:s)rt,e}}{n'_{p:s}n'_r n'_t} \quad (2)$$

where $n_p^{'}$, $n_r^{'}$, and $n_t^{'}$ were the number of students, raters, and tasks respectively while the corresponding generalizability coefficient was

$$E\rho^2 = \frac{\sigma^2(s)}{\sigma^2(s) + \sigma^2(\delta)} \tag{3}$$

Cost function and constraints The optimization procedure was constructed according to the flowchart for optimization given in Fig. 1. The purpose of optimization was to locate the optimal number of observations per facet that minimizes costs while achieving the highest possible pre-specified generalizability coefficient. The fixed design parameters were the variance components estimates from the G-study. The design variables were the sample sizes per facet. The objective function was a nonlinear cost function of $c_p \times n^{'}_p + c_t \times n^{'}_p \times n^{'}_t + c_r \times n^{'}_p \times n^{'}_t \times n^{'}_r$ where c_p is the unit cost of examining a student; c_t the unit cost of developing a task and c_r the unit cost of rating a task that is performed by a student and n'_p being the number of students in a school; $n^{'}_t$ the number of tasks and $n^{'}_r$ the number of raters. The main constraint was also a nonlinear lower-bound inequality constraint of the generalizability coefficient $(E\rho^2)$ which was varied to 0.80, 0.85, 0.90 and 0.95, respectively (Peng, Li & Wan 2012). This was in order to derive multiple optimal solutions, one for each of the four pre-set coefficients, to depict the trade-off relationship (Deb 2001). Also, the constraints that all the variables had to be greater than 1 and had to be integers were included as well. At first, the optimization procedure was conducted without any specific variable bounds. Then, acknowledging the fact that the solutions obtained might only have statistical meaning and no practical implications, variable bounds were chosen for each respective design variable to conform to realistic constraints of restricted number of students, tasks and raters. The specific values for the cost function and variable bounds are summarized in Table 1 following related research on large-scale performance assessments. The results of the two processes were compared with focus on the latter optimization procedure with constraints. In order to solve the optimization problem, mixed integer nonlinear programming with the branch and bound algorithm was employed (Bonami, Kilinç & Linderoth 2012). The optimization procedure in GT is given compactly again below:

$$\text{Minimize } 33 \times n'_p + 1 \times n'_p \times n'_t + 12 \times n'_p \times n'_t \times n'_r \tag{4}$$

$$\begin{gathered}\text{Subject to :} \\ E\rho^2 \geq g \ (g = \textit{desired reliability coefficient} = 0.80,\ 0.85,\ 0.90,\ 0.95)\end{gathered} \tag{5}$$

$$(1) \quad \textit{Without variable bounds} : 1 \leq n'_p;\ 1 \leq n'_t;\ 1 \leq n'_r \tag{6}$$

$$(2) \quad \textit{With variable bounds} : 15 \leq n'_p \leq 286;\ 2 \leq n'_t \leq 12;\ 2 \leq n'_r \leq 6 \tag{7}$$

$$n'_p, n'_t, n'_r \textit{are integers} \tag{8}$$

Table 1 Selection criteria for cost function and ranges of variable bounds

	Selection criteria
Cost for hands-on assessment (objective function)	– $22–$45 per student to administer (including materials) → Mean = $33
	– $84,000–$117,000 per task development → Divide costs among students taking the exam = $1 per task per student
	– $9–$16 per student to score → Mean = $12
Task bounds (measured using time based on standardized testing)	Minimum 2 tasks Maximum 12 tasks (= (Total test time for Grades 6∼8 = 120 minutes)/Per task time limit = 10 minutes) = 120/10)
Student bounds	Minimum 15 students Maximum 286 students (= Average number of students per grade per middle school for <Public/California/2013∼2014> = 860/3 = 286 students)
Rater bounds	Minimum 2 raters Maximum 6 raters (= (Total number of tasks)/2 = 12/2)

Note: The selection criteria came from Chingos 2013; Topol, Olson, Roeber & Hennon 2012; Stecher & Klein 1997 and National Center for Education Statistics 2015, Table 216.80

The analyses were conducted using MATLAB R2015b.

6 Results

G-study results The variance components and their relative percentages are given in Table 2. The results indicated that the (person: school) × task interaction was the most major source of measurement error, accounting for 42 % of the total variance to indicate varying degrees of difficulty felt by the students. Variance due to a student comprised 21 % of the entire variation and was twice greater than the variation among schools which suggested uncertainty about school-level performance. Furthermore, task-sampling variability was also a prevalent source of measurement error with task and task × school interaction each accounting for 6 % of the total variability; again indicating varying difficulty levels within the tasks. However, the variance component estimates for raters and rater interactions combined (excluding the residual effect) accounted for less than 2 % of the total-variability, suggesting a mediocre effect for rater-sampling variability at best (Gao et al. 1994).

Optimization Results The results of a few among many possible combinations of the facet sample sizes that produce the same pre-defined generalizability coefficients but each at a different cost and by extension, a different total number of observations (defined as $n'_p \times n'_t \times n'_r$) are given in Table 3. Such solutions are a part of the feasible region, which refers to the full set of possible values of an optimization problem that satisfies the given constraints and bounds (Bonami et al. 2012). In this case, the optimal solution within this feasible region is the solution or sample size per facet that minimizes the objective function (i.e. cheapest costs)

Table 2 Variance component estimates for the (p:s) × r × t

Source of variability	σ^2	Estimate	Percentage
s	σ^2_s	0.093	8.83
p:s	$\sigma^2_{p:s}$	0.210	19.98
r	σ^2_r	0.001	0.07
t	σ^2_t	0.061	5.79
s × r	σ^2_{sr}	0.001	0.11
s × t	σ^2_{st}	0.067	6.40
r × t	σ^2_{rt}	0.004	0.41
(p:s) × r	$\sigma^2_{(p:s)r}$	0.001	0.06
(p:s) × t	$\sigma^2_{(p:s)t}$	0.443	42.21
s × r × t	σ^2_{srt}	0.010	0.91
(p:s) × r × t	$\sigma^2_{(p:s)rt,e}$	0.160	15.24
Total number of variance components: 11			

Note: Adapted from Generalizability of large-scale performance assessments in science: Promises and problems, by Gao at el., 1994, *Applied measurement in education, 7*(4), p.332. Copyright 2009 by Taylor & Francis Ltd.. Adapted with permission.

Table 3 Various possible solutions

$E\rho^2$	n'_p	n'_t	n'_r	Cost	Total Number of observations
0.80	53	4	9	24,857	1908
0.80	33	5	5	11,154	825
0.80	16	12	2	5328	384
0.85	61	6	6	28,731	2196
0.85	32	9	4	15,168	1152
0.85	26	12	2	8658	624
0.90	155	8	10	155,155	12,400
0.90	80	10	5	51,440	4000
0.90	61	12	3	29,097	2196
0.93	210	12	8	251,370	20,160
0.93	216	12	7	227,448	18,144
0.93	238	12	5	182,070	14,280
0.95	No feasible solution				

Note: The combinations of facet sample sizes were chosen randomly from the set of feasible solutions satisfying the pre-set generalizability coefficient constraint and bound constraints (i.e. the optimization design with variable bounds)

to dominate all other solutions. This is different from the case of a one facet design where a direct linear relationship between costs, which in this case equal the total number of observations, and reliability and thus a single solution exists.

Table 4 Comparison of optimal results for the optimization design without and with variable bounds

Constraint	Without variable bounds				With variable bounds			
$E\rho^2$	n'_p	n'_t	n'_r	Cost	n'_p	n'_t	n'_r	Cost
0.80	19	10	1	3097	23	7	2	4784
0.85	29	12	1	5481	28	11	2	8624
0.90	46	19	1	12,880	67	12	2	22,311
0.95 (Max. 0.93)	109	42	1	63,111	262	12	4	162,702

The results of comparing the optimization design without and with specific bound constraints, which are organized in Table 4, showed some distinct discrepancies. For example, the latter had a restricted upper bound on the highest achievable generalizability coefficient of 0.93 within the constriants, including variable bounds, which was consistent regardless of the cost function. Also, another intuitive difference was the optimal number of observations per facet for each design as the latter design had to find a different optimal solution that was within the specified bounds. This points to the importance of the content and context of the optimization design as the optimal solution varies according to its specific settings. Moreover, it can be inferred that keeping within practical constraints is more costly than optimization procedures without variable bound constraints. Nonetheless, such constraints require careful consideration and test makers should conform to them for the results to have practical meaning. For instance, in the optimization without variable bounds, a generalizability coefficient of 0.95 would only be of statistical meaning as the sampling of 42 performance assessment tasks is not a realistic possibility.

Notwithstanding, there are similarities between the two optimization designs as well. For example, both designs showed little change in the number of raters and greatest change in the number of students where the obtained sample sizes were proportional to the magnitudes of the associated variance components. Lastly, the consistent results were that the cost of increasing reliability rose exponentially, especially in terms of optimization with variable bounds, as clearly shown by the graph in Fig. 2. For instance, while the cost of increasing the generalizability coefficient from 0.80 to 0.85 rose from $4784 to $8624, in order to get from a reliability of 0.90 to 0.93, the costs involved skyrocketed from $22,311 to $162,702. Like this, there exists a trade-off between the costs of measurement and data quality and practitioners should carefully consider whether the extra cost incurred to increase reliability is worth the investment based on various factors such as the importance of the assessment.

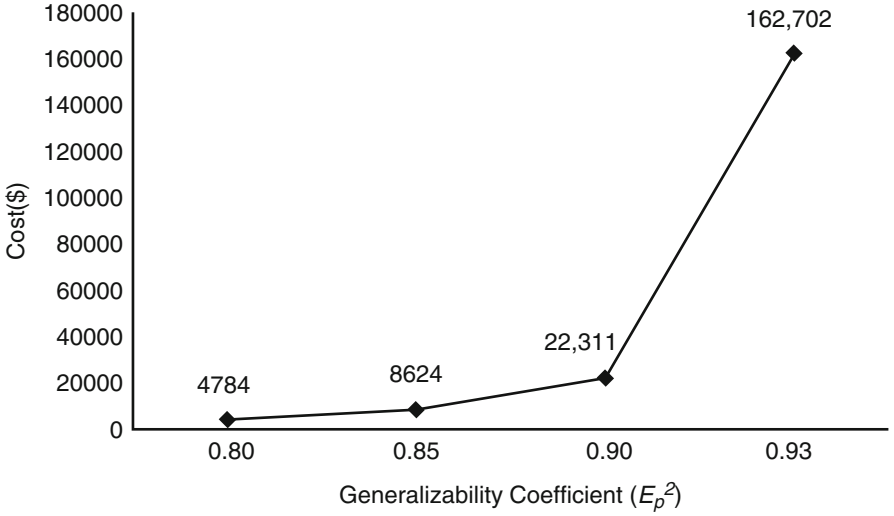


Fig. 2 Trade-off between generalizability coefficient and cost. *Note*: Trade-off curve is given for the optimization design with variable bounds

7 Conclusion

The focus of this study was to specify a cost-optimization framework incorporating GT for attaining of the most economical measurement design when various constraints are imposed, which was used for optimizing large-scale performance assessments. Succinctly, the design variables of G-study variance components were used in a subsequent D-study design common to large-scale performance assessments where minimization of cost was the objective function and both psychometric and practical constraints were enforced. The former constraint of generalizability was varied for trade-off analysis. Mixed integer nonlinear programming was used for deriving the optimal solutions of facet sample sizes.

The results of this study can be summarized as follows. First, the same level of generalizability can be achieved by multiple different combinations of sampling along the facets, proving that the relationship between the number of conditions and generalizability coefficient is not as simple as in the case of multiple choice tests with one source of error. Second, the optimization procedure was mainly influenced by the relative magnitudes of the variance components of the G-study. That is, the bigger the variance components, the larger their influences in the optimization procedure. Third, although optimization without variable bounds incurred the lowest costs, the proposed results could be unrealistic. The modified approach with variables bounds; albeit more expensive, produced more reasonable results from a practical standpoint to convey the importance of considering appropriate bound constraints. Fourth, the cost function increased non-linearly (exponentially) as the target generalizability coefficient increased. Consequently, it would be

recommended to set the GT reliability coefficient in a reasonable range after careful deliberation of the relative importance of each factor.

The findings of this study imply that among the various possible combinations of facet sample sizes arriving at the same generalizability coefficients and corresponding expenditures, the decision of which measurement procedure to choose falls ultimately into the hands of the test developers. These practitioners face a trade-off between costs and desired reliability and need to consider various practical constraints along with psychometric constraints in order for the results to be applicable to real-life assessment settings. The study shows that optimization procedures integrated with GT can promote the efficient sampling of various design factors by simultaneously taking into account such necessary constraints and depicting the trade-off relationship. Furthermore, the optimization framework described in this paper may also be applied generally and flexibly to all kinds of educational assessment settings utilizing numerous different GT study designs.

8 Limitations and Suggestions

Among the limitations and suggestions for of this study, a core consideration can be the specification of the cost function and constraints of variable bounds. The values of cost function and variable bounds of this study were chosen somewhat arbitrarily although they were based on previous research and were sufficient for demonstration purposes. There specific costs per facet and the bound constraints should be chosen carefully and befitting to the targeted assessment at hand. Also, issues related to the formulation of the cost function itself should be considered. For instance, while our cost function was defined as $c_p \times n'_p + c'_t \times n'_p \times n'_t + c'_r \times n'_p \times n'_t \times n'_r$ where there were fixed costs associated with each facet, if it were assumed that the costs per condition was more or less the same, a simple cost function of $c \times n'_p \times n'_t \times n'_r = B$ where c is the cost involved and B is the total budget would suffice (Marcoulides & Goldstein 1991). In addition, there is the possibility of the existence of various other constraints that may need to be accounted for. For instance, the practical constraint that a specific number of tasks per content area needed to be included in the assessment of interest could further be imposed. Therefore, careful attention should be given to the formulation of the optimization problem. Moreover, as stated above, there are various other possible designs that the purposed optimization framework can be applied to. Specifically, applications of the framework to nested and multivariable designs where research remains few in number may be worthwhile.

References

Antoniou, A., & Lu, W. S. (2007). *Practical optimization: Algorithms and engineering applications*. New York, NY: Springer.

Bonami, P., Kilinç, M., & Linderoth, J. (2012). Algorithms and software for convex mixed integer nonlinear programs. In *Mixed integer nonlinear programming* (pp. 1–39). New York, NY: Springer.

Brennan, R. L. (2001). *Statistics for social science and public policy: Generalizability theory*. New York, NY: Springer.

Chingos, M. M. (2013). *Standardized testing and the common core standards: You get what you pay for*. Washington, DC: Brown Center on Education Policy at Brookings. Retrieved March, 21, 2014.

Darling-Hammond, L., Newton, S. P., & Wei, R. C. (2013). Developing and assessing beginning teacher effectiveness: The potential of performance assessments. *Educational Assessment, Evaluation and Accountability, 25*(3), 179–204.

Deb, K. (2001). *Multi-objective optimization using evolutionary algorithms, 16* Hoboken, NJ: Wiley

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied measurement in education*, *7*(4), 323–342.

Jeon, M. J., Lee, G., Hwang, J. W., & Kang, S. J. (2009). Estimating reliability of school-level scores using multilevel and generalizability theory models. *Asia Pacific Education Review, 10*(2), 149–158.

Lane, S. (2010). *Performance assessment: The state of the art* (SCOPE student performance assessment series). Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Marcoulides, G. A., & Goldstein, Z. (1990). The optimization of generalizability studies with resource constraints. *Educational and Psychological Measurement, 50*(4), 761–768.

Marcoulides, G. A., & Goldstein, Z. (1991). Selecting the number of observations in multivariate measurement studies under budget constraints. *Educational and Psychological Measurement, 51*(3), 573–584.

Meyer, J. P., Liu, X., & Mashburn, A. J. (2014). A practical solution to optimizing the reliability of teaching observation measures under budget constraints. *Educational and Psychological Measurement, 74*(2), 280–291. doi: 10.1177/0013164413508774.

Parkes, J. (2000). Relationship between reliability and cost of performance assessment. *Education policy analysis archives, 8*, 16–30.

Peng, L., Li, C., & Wan, X. (2012). A framework for optimising the cost and performance of concept testing. *Journal of Marketing Management, 28*(7–8), 1000–1013.

Sanders, P. F. (1992). Alternative solutions for optimization problems in generalizability theory. *Psychometrika, 57*(3), 351–356.

Sanders, P. F., Theunissen, T. J. J. M., & Baas, S. M. (1991). Maximizing the coefficient of generalizability under the constraint of limited resources. *Psychometrika, 56*(1), 87–96.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215–232.

Stecher, B. M., & Klein, S. P. (1997). The cost of science performance assessments in large-scale testing programs. *Educational Evaluation and Policy Analysis, 19*(1), 1–14.

Topol, B., Olson, J., Roeber, E., & Hennon, P. (2012). *Getting to higher-quality assessments: Evaluating costs, benefits and investment strategies*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

U.S. Department of Education, National Center for Education Statistics. (2015). Table 216.80: Public secondary schools, by grade span, average school enrollment, and state or jurisdiction: 2013–14. In U.S. Department of Education, National Center for Education Statistics (Ed.), *Digest of Education Statistics* (2015 ed.). Retrieved from https://nces.ed.gov/programs/digest/d15/tables/dt15_216.80.asp.

Woodward, J. A., & Joe, G. W. (1973). Maximizing the coefficient of generalizability in multi-facet decision studies. *Psychometrika, 38*(2), 173–181.

A Confirmatory Factor Model for the Investigation of Cognitive Data Showing a Ceiling Effect: An Example

Karl Schweizer

Abstract A method for addressing the consequences of ceiling effects on model misfit in confirmatory factor analysis of cognitive data is proposed. This method focuses on the reduction of variance as a major ingredient of the ceiling effect. The model of the covariance matrix is modified in such a way that it reflects the impact of the ceiling effect on variances and covariances. The method applies to models including theory-based constraints of factor loadings for investigating cognitive data. The effectiveness of the method is demonstrated in data collected by means of a measure of working memory capacity. The application of the method in combination with a confirmatory factor model that assumes working memory capacity as the major source of performance yields the expected increase in the degree of model fit.

Keywords Ceiling effect • Confirmatory factor analysis • Model of the covariance matrix • Cognitive data • Constrained factor loadings

1 Introduction

Items of measures capturing cognitive processes that exhibit a high degree of efficiency, or so-called speeded tests, are likely to yield scores showing a ceiling effect. A major characteristic of this effect is the reduction of the variance of scores. This reduction of variance can be a source of model misfit. The modeling of the reduction in variance is the core of the proposed method that is expected to mitigate model misfit.

K. Schweizer (✉)
Department of Psychology, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 6, Frankfurt a. M. 60323, Germany
e-mail: K.Schweizer@psych.uni-frankfurt.de

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_14

1.1 The Ceiling Effect

The observation of a large number of participants reaching maximum scores or achieving scores close to the maximum score in psychological assessment is referred to as a ceiling effect (Hessling, Traxel & Schmidt, 2004; Vogt 2005). The result is a distortion of the data distribution especially if normality of the data distribution is expected. The ceiling effect is associated with a reduction of variance due to the limitation of the range of possible scores. However, because of the lack of a criterion, it is not possible to clearly distinguish between a smaller than expected variance occurring at random and a manifestation of the ceiling effect.

Textbooks on test construction recommend limiting the range of difficulties of items (e.g., Allen & Yen 2001). This limitation means the avoidance of extremely easy items and of manifestations of the ceiling effect. However, the investigation of basic cognitive processes and basic properties of human information processing may lead to departures from this recommendation because the relevant processes may show a high degree of automation and can be performed so efficiently that a high degree of easiness can hardly be avoided. A well-known example of a measure focusing on such processes is the Rapid Visual Information Processing task (Wesnes & Warburton 1983) that is still in use although it was presented quite a time ago. Furthermore, in longitudinal research the ceiling effect can be the result of developmental processes (Wang, Zhang, McArdle & Salthouse 2009). Whereas in young participants the scores may show a suitable distribution, after some years the same participants may produce scores that reflect a ceiling effect.

In order to avoid negative consequences of the ceiling effect in statistical investigations, the following methods are employed: listwise deletion alone, listwise deletion, whereby data showing the ceiling effect are treated as missing data especially through the use of Tobit regression (Tobit 1958), as is proposed by Muthen (1989). However, these methods are not without disadvantages. Listwise deletion means a decrease of the sample size, and there is the danger of the removal of a homogeneous subset of participants sharing a specific property. The treatment as missing data in the sense of replacement by the means of an estimation algorithm cannot really be helpful because of the limitation to the range of possible scores and the necessity to distinguish between acceptable maximum scores and maximum scores to be considered as missing data. Furthermore, the use of Tobit regression requires the assumption that the measurements of all participants with maximum scores show no error, and additionally there is the observation that this regression method biases factor loadings downwards (Van den Oord & Van der Ark 1997).

The reduction of variance that is considered as the major characteristic of the ceiling effect by many authors (e.g., Hessling et al. 2004; Uttl 2005; Vogt 2005) is very disadvantageous for statistical investigations. It can mean a major deviation of the variance of the variable showing the ceiling effect from the variances of other variables assumed to depend on the same source. Some statistical methods, such as analysis of covariance (Jöreskog 1970), include a model of measurement that relates a set of observed variables to one or a few latent sources. This model of

measurement implicitly assumes that there is common latent variance and that the differences between the additional observed variances are due to random variation. In applications of such methods the ceiling effect may result in a rather small estimate of the common latent variance, and it may even happen that negative error variance is observed for the variable showing the ceiling effect.

Because of the importance of unbiased variances for statistical investigations, the reasoning in the following sections focuses on the manifestation of the ceiling effect as the reduction of variance.

1.2 The Representation of the Ceiling Effect

The first step in the development of a method for addressing the influence of ceiling effects on model misfit focuses on the reduction of variance. In the absence of a cut-off for the degree of reduction, all deviations of the observed variance from the full variance of the reference variable characterized by the absence of the ceiling effect are considered. Assume the random variables Y_i ($i = 1, \ldots, N_S$) and Y_u representing cognitive scores with reduced and full variances $\text{var}(Y_i)$ and $\text{var}(Y_u)$ respectively, and weight w_i ($0 < w_i \leq 1$) for representing the relationship between the reduced and full variances such that

$$\text{var}(Y_i) = w_i^2 \, \text{var}(Y_u) . \tag{1}$$

In cognitive research Y_i and Y_u are usually scores that are obtained by summing the binary outcomes {0,1} of a large number of trials. These trials are considered as repeated measurements, and the scores are treated as continuous variables in statistical investigations.

In the next step the type of distribution that the scores are assumed to follow needs to be specified. The binomial distribution characterized as $Bin(n, p)$ with n and p as the number of binary events that is in this case the number of binary items contributing to the score and the probability of a correct response [$p = \Pr(X = 1)$] respectively is selected for three reasons: it enables the modeling of frequency distributions that are similar to actual distributions of cognitive scores; the full variance is always given as the variance associated with the probability of .5, and only two parameters need to be considered.

The selection of this distribution implies a switch from Y_i and Y_u to X_{ij} and X_{uj} ($i = 1, \ldots, N_S$, $j = 1, \ldots, N_E$) that are assumed to constitute Y_i and Y_u where i refers to scores and j to the items included in scores. Furthermore, assuming that the binary events giving rise to Y_i and Y_u show the same probability, $Bin(n, p)$ is actually $Bin(1, p)^n$, i.e. a vector of random variables showing binomial distributions with $n = 1$. The clarification regarding the type of distribution leads to the following equation regarding Y_i that replaces Eq. 1:

$$N_{\mathrm{E}} \Pr(X_i = 1)\,[1 - \Pr(X_i = 1)] = w_i^2 N_{\mathrm{E}} \Pr(X_{\mathrm{u}} = 1)\,[1 - \Pr(X_{\mathrm{u}} = 1)] \quad (2)$$

where $\Pr(X_i = 1) = \Pr(X_{i1} = 1) = \ldots = \Pr(X_{iN\mathrm{E}} = 1)$ and $\Pr(X_{\mathrm{u}} = 1) = \Pr(X_{\mathrm{u}1} = 1) = \ldots = \Pr(X_{\mathrm{u}N\mathrm{E}} = 1) = .5$ since equality of the probabilities is assumed. Furthermore, since N_{E} is a factor of each one of the products serving as left- and right-hand parts of this Equation, it can be removed:

$$\Pr(X_i = 1)\,[1 - \Pr(X_i = 1)] = w_i^2 \Pr(X_{\mathrm{u}} = 1)\,[1 - \Pr(X_{\mathrm{u}} = 1)]\,. \quad (3)$$

After the reordering of the components and the isolation of weight w_i, a definition of w_i is available:

$$w_i = \sqrt{\Pr(X_i = 1)\,[1 - \Pr(X_i = 1)]\,/0.25} \quad (4)$$

where the denominator that is 0.25 is the variance of the reference variable X_{u} var($\mathrm{X_u}$) since a main characteristic of the reference variable is that $\Pr(X_{\mathrm{u}} = 1) = .5$. This weight can be expected to do well if the assumed underlying source contributes to each random variable Y_i approximately equally.

In contrast, if different values of i signify that different variances are expected, the variance of the reference variable X_{u} has to be adapted to the altered expectation. This means that each i ($i = 1, \ldots, N_{\mathrm{S}}$) is associated with another reference variable $\mathrm{X_{u}}_i$ and another variance var($\mathrm{X_{u}}_i$). In Eq. 5 the denominator is modified accordingly:

$$w_i = \sqrt{\Pr(X_i = 1)\,[1\text{-}\ \Pr(X_i = 1)]\,/\,\mathrm{var}(X_{\mathrm{u}i})}. \quad (5)$$

In an application the variance of the reference variable must be defined in considering the assumptions of cognitive information processing.

1.3 The Integration of the Representation of the Ceiling Effect into the Model of the Covariance Matrix

This section describes how the representation of the ceiling effect is integrated into the model of the covariance matrix. The model-implied $N_{\mathrm{S}} \times N_{\mathrm{S}}$ covariance matrix Σ is modified in such a way that it not only reflects the assumed cognitive sources of responding but also the ceiling effect. Normally it represents the assumed cognitive sources of responding only. The modification is expected to eliminate the part of the difference between Σ and the $N_{\mathrm{S}} \times N_{\mathrm{S}}$ empirical covariance matrix $\mathbf{S}$ that is due to the ceiling effect.

The model of the $N_{\mathrm{S}} \times N_{\mathrm{S}}$ covariance matrix Σ is given by

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}. \quad (6)$$

with Λ as the $N_S \times N_V$ matrix of factor loadings, $\mathbf{\Phi}$ as the $N_V \times N_V$ matrix of the variances and covariances of the N_V latent variables and Θ as the $N_S \times N_S$ diagonal matrix of error variances (Jöreskog 1970). This model includes systematic variances and covariances ($\Lambda\Phi\Lambda'$) on one hand and error variances (Θ) on the other hand that are addressed as the systematic and error parts in this chapter.

The weights bear on the systematic part of the model since the ceiling effect is assumed not to be at random but reflects an undesirable property of an item or score. The weights are assumed to vary between zero and one such that in the case of a weight of one nothing is changed. Therefore, it is not necessary to treat the absence of the ceiling effect as a special case. In order to assure that each weight refers to the corresponding random variable, the weights w_i ($i = 1, \ldots, N_S$) are integrated into the $N_S \times N_S$ diagonal matrix of weights $\mathbf{W}$:

$$\mathbf{W} = \begin{bmatrix} w_1 & & 0 \\ & w_i & \\ 0 & & w_m \end{bmatrix}. \tag{7}$$

This weight matrix modifies the systematic part of the model according to Eq. 6 so that it reflects the impact of the ceiling effect:

$$\mathbf{\Sigma} = \mathbf{W}\left(\mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}'\right)\mathbf{W}' + \mathbf{\Theta}. \tag{8}$$

This way of integrating the representation of the ceiling effect into the model of the covariance matrix is in line with the assumption that the ceiling effect occurs during the assessment process as a distortion that shrinks the variance. In doing so, it avoids assuming error-free measurement, as in the Tobit model (Tobit 1958; Van den Oord & Van der Ark 1997). Only a partial modification of the implicit model of measurement may be necessary. Such a measurement model has to include several components referring to the various scores contributing to the scale that is investigated. Only the components associated with scores showing a ceiling effect should be reflected in the corresponding weights.

1.4 Models with Constrained Factor Loadings

Negative consequences of ceiling effects on model fit are especially likely if the factor loadings are constrained to represent specific structural assumptions, as the compensation for large deviations of the variances may occur in the freely estimated factor loadings. Such constraints mean fixed discriminability of the model of measurement; they characterize the tau-equivalent model (Lord & Novick 1968), the Rasch model (Rasch 1960), the corresponding one-parameter item-response

model (Birnbaum 1968) and the linear logistic test model (Kubinger 2009). The concentration on models with constrained factor loadings is made apparent by replacing the $N_S \times N_V$ matrix of factor loadings Λ in Eq. 8 by the $N_S \times N_V$ matrix of constraints B:

$$\boldsymbol{\Sigma} = \mathbf{W}\left(\mathbf{B}\boldsymbol{\Phi}\mathbf{B}'\right)\mathbf{W}' + \boldsymbol{\Theta}. \tag{9}$$

Since both **W** and **B** include no parameters that need to be estimated, these matrices are rearranged and separated from Φ of which some or all elements are estimated in confirmatory factor analysis:

$$\boldsymbol{\Sigma} = (\mathbf{WB})\,\boldsymbol{\Phi}\,(\mathbf{WB})' + \boldsymbol{\Theta}. \tag{10}$$

The numbers included in one column of matrix **B** reflect a specific hypothesis regarding structure. In the simplest case the hypothesis is reflected by one column including equal-sized numbers. Different numbers included in the same column of **B** suggest that the common source accounts for different amounts of variance of the corresponding manifest variables.

2 A Real Data

The Exchange Test (Schweizer 1996) was introduced as a measure of working memory capacity that demanded the exchange of the positions of neighboring elements of a series of symbols in order to establish equivalence of the sequences of this series and of a second series of symbols. It included five treatment levels and 12 trials per treatment level. The outcome of a trial was the correctness of the number of necessary exchanges to be determined by the participant. Different numbers of necessary exchanges characterized the trials of the five levels (1, 2, 3, 4, 5 for the first to fifth levels).

Because of the effect of an exchange operation on the load in cognitive processing, each increase in level was expected to give rise to an increase in variance. Therefore, the matrix of constraints $\mathbf{B}_{WM}$ reflecting the load on working memory had to include digits according to the numbers of necessary exchanges:

$$\mathbf{B}_{WM} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}. \tag{11}$$

Since performance according to the demands of a cognitive measure usually involved contributions of additional processes besides the core processes, as for

example regarding the motor response, additional systematic variance also needed to be represented. Since these auxiliary processes (AP) contributed in the same way to every treatment level, the same number was assigned to the entries of an extra column of matrix $\mathbf{B}_{\mathrm{WM+AP}}$:

$$\mathbf{B}_{\mathrm{WM+AP}} = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \\ 5 & 1 \end{bmatrix}. \tag{12}$$

The processes represented by the two latent variables associated with the two columns are very different and, therefore, expected to be independent of each other. Whereas the exchange processes are assumed to reflect working memory capacity, the auxiliary processes should be unrelated to it. Therefore, these latent variables are not allowed to correlate with each other.

Since there were different theory-based expectations for the treatment levels, as is obvious from Eq. 11, it appeared that the variance of the reference variable had to be adapted to these expectations. Therefore, in the application the weights were computed according to Eq. 5 in addition to weights according to Eq. 4. The variance of the reference variable included in the denominator was modified accordingly:

$$\mathrm{var}\,(X_{\mathrm{u}i}) = \Pr\,(X_{\mathrm{u}i} = 1)\,[1 - \Pr\,(X_{\mathrm{u}i} = 1)] \tag{13}$$

where i gave the treatment level. In analogy to the numbers included in the vector of Eq. 11, equal-sized distances between the probabilities were selected so that there was a linear decrease from .9 to .5 ($\Pr(X_i = 1) = .9, .8, .7, .6, .5$; $i = 1, \ldots, 5$).

The computation of weights led to the numbers presented in Table 1. The first column of this Table includes the probabilities computed from the binary outcomes of completing the items of each level. The second column provides the weights that were computed by means of Eq. 4 and the third column weights achieved by means of Eqs. 5 and 13. These weights were assigned to the main diagonal of the matrix of Eq. 7.

The data of 345 university students were investigated by means of LISREL (Jöreskog & Sörbom 2006) with the variance-covariance matrix as input. The estimation occurred by means of the maximum likelihood (ML) and robust maximum likelihood (RML) methods. Only ML was applied in combination with weights since weights did not lead to an improvement in model fit in combination with RML.

The model included two latent variables and the five treatment levels of Exchange Test as manifest variables. It is referred to as the *basic model*. The model with weights is denoted the *weighted model*. There were several versions including different numbers of weights. Furthermore, there were two types of weights. The first type included weights according to Eq. 4 (unspecific: US) and the second type according to Eq. 5 and 13 (WM related: WM).

Table 1 Probabilities characterizing the trials of the five treatment levels and corresponding Weights computed according to Eqs. 4, 5 and 13

Treatment level	Probability	Weights to be used in the absence of specific expectations	Weights for WM-based expectations
1	.974	0.318	0.530
2	.891	0.623	0.779
3	.767	0.645	0.922
4	.660	0.947	0.966
5	.500	1	1

Table 2 Fit results for models with and without adjustment for preventing poor model fit due to the ceiling effect in data collected by means of exchange test

Model and characteristics	χ^2	df	RMSEA(90 %)	SRMR	CFI	TLI	GFI
Basic model (ML)	29.48	8	.088(.06, .12)	.077	.94	.93	.97
Basic model (RML)	21.86	8	.071(.04, .11)	.062	.98	.97	.99
Weighted model with ...							
... level 1 US adjusted	33.33	8	.096(.06, .13)	.092	.93	.92	.96
... levels 1–2 US adjusted	18.14	8	.061(.02, .10)	.056	.97	.97	.98
... levels 1–3 US adjusted	24.26	8	.077(.04, .14)	.071	.96	.95	.97
... levels 1–4 US adjusted	20.55	8	.068(.03, .10)	.063	.97	.96	.98
... level 1 WM adjusted	25.59	8	.080(.05, .12)	.075	.95	.94	.97
... levels 1–2 WM adjusted	19.35	8	.064(.03, .10)	.058	.97	.96	.98
... levels 1–3 WM adjusted	22.97	8	.074(.04, .11)	.068	.96	.95	.97
... levels 1–4 WM adjusted	21.07	8	.069(.03, .10)	.064	.97	.96	.98

The fit results regarding chi-square, RMSEA (root mean square error of approximation), SRMR (standardized root mean square residual), CFI (comparative fit index), NNFI (non-normed fit index) and GFI (goodness of fit index) for the basic and weighted models are included in Table 2.

The criteria for evaluating these fit statistics were from the study by Hu and Bentler (1999). They suggested .06, .08, .95, .95 and .90 as upper or lower limits for RMSEA, SRMR, CFI, NNFI and GFI in corresponding order. The first row gives the results for the basic model obtained by means of the maximum likelihood estimation method. Most of the fit statistics indicated an acceptable fit but not a good one. The majority of the statistics of the second row obtained by means of robust maximum likelihood estimation indicated a good fit. However there was still the RMSEA result that was not good and the ratio of chi-square to degrees of freedom was also not good.

The results reported in the following rows were achieved for models that included weights computed in considering one of two types of weights (unspecific: US, WM related: WM). Furthermore, the rows differed according to the number of considered weights. Four of the eight combinations of type of weight and number of weights

led to an improved RMSEA result (. . . : levels 1–2 US adjusted, . . . : levels 1–4 US adjusted, . . . levels 1–2 WM adjusted, . . . levels 1–4 WM adjusted).

The overall best model fit according to chi-square, RMSEA and SRMR was achieved for the version characterized by weights added to the levels 1 and 2 and computed according to Eq. 4 (US). Only this version yielded an RMSEA value close to the cut-off provided by Hu and Bentler (1999). The basic model in combination with robust maximum likelihood estimation yielded CFI and GFI values that slightly surmounted the values observed for the best version of the weighted model. However, the CFI and GFI observed for the best version of the weighted model already indicated good model fit. The difference in chi-square between the basic and best weighted model was 11.34 when there was customary maximum likelihood estimation. In the case of robust maximum likelihood estimation it was 3.72.

Since the Exchange Test was a measure of working memory capacity, it was expected that the corresponding latent variable accounted for most of the latent variance. In the best-fitting weighted confirmatory factor model the latent variances were 0.21 ($t = 4.72$) and 2.00 ($t = 4.15$) for the latent variables representing working memory capacity and auxiliary processes in corresponding order. The corresponding scaled variances (Schweizer 2011) were 2.18 and 1.40. These numbers signified that 60.9 % of the variance at the latent level was due to working memory capacity and the remaining 39.1 % due to auxiliary processes.

3 Conclusions

The ceiling effect is the consequence of the use of one or more items contributing to a score that are too easy for the sample. This effect can be avoided by carefully selecting items according to the recommendations for test construction (e.g., Allen & Yen 2001) and representative sampling. However, there are situations that demand the assessment of processes that can be performed almost without effort. Furthermore, speeded testing may also likely entail a ceiling effect. Because of these special situations that are quite common in the area of cognitive research, it is necessary to preserve as much information as possible as opposed to simply eliminating variables displaying ceiling effects.

The definition of the ceiling effect as the observation of a large number of participants reaching the maximum score suggests a distortion of the distribution of scores, and it is the distortion of scores that leads to the reduction of variance. As has been demonstrated in empirical datasets expected to follow the normal distribution, deviations from the normality assumption are more often present than not (Micceri 1989), and there are many ways of deviating from normality. However, irrespective of the kind of deviation, what counts most with respect to statistical investigations is the reduction of variance.

If there is an estimate of the original variance that can serve as expected variance, taking the ratio between the observed and expected variances as a ratio between the reduced and original variances is valuable information that can be used for

computing weights which reflect the extent of the ceiling effect. The weight-based method for addressing model misfit due to the ceiling effect is very specific. It modifies the components of the model of measurement that are affected by the ceiling effect, and the modification reflects the extent of the distortion of the data. It is this specificity that distinguishes this method from some other methods that are recommended for dealing with the ceiling effect.

The selection of the binomial distribution as a major ingredient of the method may appear to be self-evident since binary events are basic to the data. However, the selection of this distribution is an assumption, and there may be situations where it does not apply. In this regard it appears to be important that there is invariance regarding subsamples (Vandenberg & Lance 2000) since differing subsamples can mean a distortion of the data distribution.

The use of weights led to the expected good model fit according to virtually all fit statistics although there was only customary maximum likelihood estimation. Robust estimation (Satorra & Bentler 1994) was additionally considered and found to improve model fit over the model fit observed without weights for preventing impairment due to the ceiling effect. However, an attempt to combine the proposed method with robust estimation was not successful and, therefore, is not reported. Further research is necessary in order to enable the use of robust estimation and weights for addressing model misfit due to the ceiling effect.

References

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory* (2nd ed.). Monterey CA: Brooks/Cole.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Hessling, R. M., Traxel, N. M., & Schmidt, T. J. (2004). Ceiling effect. In M. S. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods* (p. 107). Thousand Oaks: Sage.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. doi:10.1080/10705519909540118.

Jöreskog, K. G. (1970). A general method for analysis of covariance structure. *Biometrika, 57*, 239–257. doi:10.2307/2334833.

Jöreskog, K. G., & Sörbom, D. (2006). *LISREL 8.80.*. Lincolnwood, IL: Scientific Software International.

Kubinger, K. D. (2009). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement, 69*, 232–244. doi:10.1177/0013164408322021.

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

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156–166. doi:10.1037/0033-2909.105.1.156.

Muthen, B. (1989). TOBIT factor analysis. *British Journal of Mathematical and Statistical Psychology, 42*, 241–250. doi:10.1111/j.2044-8317.1989.tb00913.x.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Satorra, A., & Bentler, P. M. (1994). Corrections to the test statistics and standard errors on covariance structure analysis. In A. von Eye & C. C. Glogg (Eds.), *Latent variable analysis* (pp. 399–419). Thousand Oaks, CA: Sage.

Schweizer, K. (1996). The speed-accuracy transition due to task complexity. *Intelligence, 22*, 115–128. doi:10.1016/S0160-2896(96)90012-4.

Schweizer, K. (2011). Scaling variances of latent variables by standardizing loadings: Applications to working memory and the position effect. *Multivariate Behavioral Research, 46*, 938–955. doi:10.1080/00273171.2011.625312.

Tobit, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica, 26*, 24–36. doi:10.2307/1907382.

Uttl, B. (2005). Measurements of individual differences: Lessons from memory assessment in research and clinical practice. *Psychological Science, 16*, 460–467. doi:10.1111/j.0956-7976.2005.01557.x.

Van den Oord, E. J. C. G., & Van der Ark, L. A. (1997). A note on the use of Tobit approach for tests scores with floor or ceiling effects. *British Journal of Mathematical and Statistical Psychology, 50*, 351–364. doi:10.1111/j.2044-8317.1997.tb01150.x.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organisational research. *Organisational Research Methods, 16*, 4–70.

Vogt, W. P. (2005). *Dictionary of statistics & methodology: A nontechnical guide for the social sciences* (3rd ed.). Thousand Oaks: Sage.

Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2009). Investigating ceiling effects in longitudinal data analysis. *Multivariate Behavioral Research, 43*, 476–496. doi:10.1080%2F00273170802285941.

Wesnes, K., & Warburton, D. M. (1983). The effect of smoking on rapid visual information processing performance. *Neuropsychobiology, 9*, 223–229. doi:10.1016/j.neurobiolaging.2014.01.145.

The Goodness of Sample Loadings of Principal Component Analysis in Approximating to Factor Loadings with High Dimensional Data

Lu Liang, Kentaro Hayashi, and Ke-Hai Yuan

Abstract Guttman (Psychometrika 21 273–286:1956) showed that the loadings of factor analysis (FA) and those of principal component analysis (PCA) approach each other as the number of variables p goes to infinity. Because the computation for PCA is simpler than FA, PCA can be used as an approximation for FA when p is large. However, another side of the coin is that as p increases, non-consistency might become an issue. Therefore, it is necessary to simultaneously consider the closeness between the estimated FA and the estimated PCA loadings as well as the closeness between the estimated and the population FA loadings. Using Monte Carlo simulation, this article studies the behavior of three kinds of closeness under high-dimensional conditions: (1) between the estimated FA and the estimated PCA loadings, (2) between the estimated FA and the population FA loadings, and (3) between the estimated PCA and the population FA loadings. To deal with high-dimensionality, a ridge method proposed by Yuan and Chan (Computational Statistics and Data Analysis 52:4842–4828, 2008) is employed. As a measure for closeness, the average canonical correlation (CC) between two loading matrices and its Fisher-z transformation are employed. Results indicate that the Fisher-z transformed average CC between the estimated FA and the estimated PCA loadings is larger than that between the estimated FA and the population FA loadings as well as that between the estimated PCA and the population FA loadings. It is concluded that, under high-dimensional conditions, the closeness between the estimated FA and PCA loadings is easier to achieve than that between the estimated and the population FA loadings and also that between the estimated PCA and the population FA loadings.

L. Liang • K. Hayashi (✉)
Department of Psychology, University of Hawaii at Manoa, 2530 Dole Street, Sakamaki C400, Honolulu, HI 96822, USA
e-mail: lianglu@hawaii.edu; hayashik@hawaii.edu

K.-H. Yuan
Department of Psychology, University of Notre Dame, 123A Haggar Hall, Notre Dame, IN 46556, USA
e-mail: kyuan@nd.edu

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_15

Keywords Canonical correlation • Factor indeterminacy • Fisher-z transformation • Guttman condition • Large p small N • Ridge factor analysis

1 Introduction

Factor analysis (FA) and principal component analysis (PCA) are frequently used multivariate statistical methods for data reduction. In FA (Anderson 2003; Lawley & Maxwell 1971), the mean-centered vector of the observed variables $\boldsymbol{y}$ $(p \times 1)$ is linearly related to a vector of latent factors $\boldsymbol{f}$ $(m \times 1)$ as $\boldsymbol{y} = \boldsymbol{\Lambda f} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\Lambda}$ $(p \times m$, with $p > m)$ is a matrix of factor loadings, and $\boldsymbol{\varepsilon}$ $(p \times 1)$ is a vector of errors. For the orthogonal factor model, three assumptions are typically imposed: (1) $\boldsymbol{f} \sim N_m(\boldsymbol{0}, \boldsymbol{I}_m)$; (2) $\boldsymbol{\varepsilon} \sim N_p(\boldsymbol{0}, \boldsymbol{\Psi})$, with $\boldsymbol{\Psi}$ diagonal; (3) $Cov(\boldsymbol{f}, \boldsymbol{\varepsilon}) = \boldsymbol{0}$. Then, under the three assumptions, the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{y}$ is given by $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$.

Let $\boldsymbol{\Omega}$ $(m \times m)$ be the diagonal matrix whose elements are the first m largest eigenvalues of $\boldsymbol{\Sigma}$, and $\boldsymbol{\Lambda}^+$ $(p \times m)$ be the matrix whose columns are the standardized eigenvectors corresponding to $\boldsymbol{\Omega}$. Then the first m principal components (PCs) (c.f., Anderson 1963, 2003) are obtained as $\boldsymbol{f}^* = \boldsymbol{\Lambda}^{+'}\boldsymbol{y}$. Clearly, the PCs are uncorrelated with each other, with a covariance matrix $\boldsymbol{\Omega}$. When m is properly chosen, there exists $\boldsymbol{\Sigma} \approx \boldsymbol{\Lambda}^+\boldsymbol{\Omega}\boldsymbol{\Lambda}^{+'} = \boldsymbol{\Lambda}^*\boldsymbol{\Lambda}^{*'}$, where $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}^+\boldsymbol{\Omega}^{1/2}$ $(p \times m)$ is the matrix of PCA loadings, with $\boldsymbol{\Omega}^{1/2}$ being the diagonal matrix whose elements are the square root of those in $\boldsymbol{\Omega}$.

It has been well-known that FA and PCA often yield approximately the same loading matrices $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\Lambda}}^*$, respectively. Conditions under which the two matrices are close to each other have been studied extensively (Bentler & Kano 1990; Krijnen 2006; Schneeweiss 1997; Schneeweiss & Mathes 1995). The single most well-known condition for closeness at the population level was identified by Guttman (1956), which requires that $p \to \infty$ while $m/p \to 0$. In words, as p increases faster than m, PCA gives a close approximation to FA. Because the computation for the estimates of PCA loadings is much simpler than that for the estimates of FA loadings in that the former is just an eigenvalue-eigenvector decomposition of the sample covariance matrix, it is attractive if PCA can be used as an approximation for FA when p is large. Recently, with the advancement of computing technology, analysis of high-dimensional data with large p is becoming easier, yet the amount of computation is still demanding. Thus, using PCA as an approximation for FA under high-dimensionality is a viable alternative if they yield similar results.

However, here, it is important to note that non-consistency might become an issue as p increases. In fact, Johnstone and Lu (2004, 2009) showed that $\hat{\boldsymbol{\Lambda}}^*$ is a consistent estimate of $\boldsymbol{\Lambda}^*$ in a single-component PCA if and only if $p/N \to 0$. An extension of Johnstone and Lu (2004, 2009) to a multi-component PCA was given by Paul (2007). Likewise, under a high-dimensional setting and when both p and N approach infinity, Bai and Li (2012) studied FA and PCA, and showed that

their loading estimates converge to the same asymptotic normal distribution, where the assumption of $\sqrt{p}/N \to 0$ along with $p, N \to \infty$ is needed. Obviously, the condition of $p/N \to 0$ in Johnstone and Lu (2004, 2009) implies $\sqrt{p}/N \to 0$.

These studies suggest that the closeness between the estimates of FA loadings and the estimates of PCA loadings is not sufficient for trusting the estimates of PCA loadings as a good approximation for the population FA loadings. We also need to consider the closeness between the estimated FA loadings and their population FA loadings. If the discrepancy is large between the estimated and the population FA loadings, then the closeness between the estimates of FA loadings and the estimates of PCA loadings may have little inferential value.

Although some authors call only data with $p > N$ as high-dimensional (e.g., Pourahmadi 2013), we do not require this assumption to accommodate typical social science data. Also, we do not consider a covariance matrix that has many zero entries, called sparsity. Our approach is to study FA and PCA with high-dimensional data in the classical framework, except that we employ the ridge regression approach to the covariance matrix introduced by Yuan and Chan (2008) and Yuan (2013).

2 Purpose of Study

We examine the relationship among three different kinds of closeness under high-dimensionality, that is, (1) the closeness between the estimates of the FA loading matrix and the estimates of the PCA loading matrix, (2) the closeness between the estimates of the FA loading matrix and the population FA loading matrix, and (3) the closeness between the estimates of the PCA loading matrix and the population FA loading matrix.

At the population level, Guttman (1956) showed that as $p \to \infty$ with $m/p \to 0$, the FA loadings and PCA loadings converge. Liang, Hayashi, and Yuan (2015) as well as many others have confirmed that on sample estimates through simulations. For a fixed number of variables p, it is known that the estimates of factor loadings converge to their population counterparts at the rate of $O\left(N^{-1/2}\right)$ (Anderson 2003; Lawley & Maxwell 1971). Regarding the case where $p, N \to \infty$, Johnstone and Lu (2004, 2009) showed that the consistency for the PCA loadings holds if and only if $p/N \to 0$, which implies $p \to \infty$ at a slower rate than $N \to \infty$. However, beyond that, little is known regarding how the two kinds of closeness are inter-connected. Thus, the main goal of our work is to investigate the relationship among the three kinds of closeness through a simulation study.

Our emphasis is on high dimensional settings where p is relatively close to N. To the best of our knowledge, there have not been any simulation studies systematically examining the relationships simultaneously among the three kinds of closeness to date.

3 Simulation Conditions

The population factor loading matrix in our study is of the following form: $\mathbf{\Lambda} = \mathbf{1}_q \otimes \mathbf{\Lambda}_{12}$, where $\mathbf{1}_q$ is the column vector of q 1's,

$$\mathbf{\Lambda}_{12}{}' = \begin{pmatrix} \lambda_{11} & \lambda_{21} & \lambda_{31} & \lambda_{41} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_{52} & \lambda_{62} & \lambda_{72} & \lambda_{82} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{93} & \lambda_{10,3} & \lambda_{11,3} & \lambda_{12,3} \end{pmatrix},$$

and $\otimes$ is the Kronecker product. Thus, there are $m = 3$ factors and the number of indicators for each factor are multiplications of 4. Two conditions of population loadings are employed: (1) *equal loadings*: $\lambda_{ij} = 0.8$ for every non-zero factor loading, and (2) *unequal loadings*: $\lambda_{11} = \lambda_{21} = \lambda_{52} = \lambda_{62} = \lambda_{93} = \lambda_{10,3} = 0.8$, $\lambda_{31} = \lambda_{72} = \lambda_{11,3} = 0.75$, $\lambda_{41} = \lambda_{82} = \lambda_{12,3} = 0.7$. The numbers of observed variables are multiples of 12: $p = 12q$, where $q = 2^k$ ($k = 1, 2, \ldots, 7$) with $p/N < 0.5$. The detailed information on p and N is given in Table 1, where p ranges from 12 to 1536, N ranges from 200 to 3200, and when q is greater than 1, we stack the structure of $\mathbf{\Lambda}_{12}$ vertically. The factors are orthogonal so that the population covariance structures are of the form: $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$, where the diagonal elements of $\mathbf{\Sigma}$ are all 1's. As a result, (1) corresponds to equal unique variances (the Equal Psi covariance structure) and (2) corresponds to unequal unique variances (the Unequal Psi covariance structure) in the population. Note that FA with equal unique variances can be considered as a variant of PCA (Johnstone & Lu 2004, 2009). For one factor model with equal unique variances, Bentler and Kano (1990) showed that the PCA loading vector can be expressed as a function of the FA loading vector. On the other hand, FA with unequal unique variances can only be approximated by PCA. There does not exist any analytical formula that connects the PCA loadings as a function of the FA loadings.

Let $\boldsymbol{S}$ be the sample covariance matrix, and we perform FA on $\boldsymbol{S}_a = \boldsymbol{S} + a\boldsymbol{I}_p$, referred to as ridge FA, where $\boldsymbol{I}_p$ ($p \times p$) is the identity matrix and a is a tuning parameter. In the analysis, we let $a = p/N$ as was recommended in Yuan and Chan

Table 1 Combination of (p, N) pairs in the simulation study

p	N	200	400	566	800	1131	1600	2236	3200
12		X	X	X	X	X	X	X	X
24		X	X	X	X	X	X	X	X
48		X	X	X	X	X	X	X	X
96		X	X	X	X	X	X	X	X
192			X	X	X	X	X	X	X
384					X	X	X	X	X
768							X	X	X
1536									X

Note: "X" stands for the (p, N) pairs used in our simulation study. We employed the (p, N) pairs only if $p/N < 0.5$

(2008) and Yuan (2013), which led to more accurate estimates of the factor loadings than performing FA on $\boldsymbol{S}$. No attempt to identify an optimal tuning parameter is made. We perform PCA on $\boldsymbol{S}$, not on $\boldsymbol{S}_a$.

For each condition of N, p and $\boldsymbol{\Sigma}$, we performed 100 replications of samples from the multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. For each replication, to obtain the maximum likelihood estimates of FA loadings, we employed the "factanal" function in the R language (see., e.g., Beaujean 2013) and modified it to fit our simulation purpose. We used the default convergence criterion set by the "optim" function. For PCA, we simply used the "eigen" function in R to find the eigenvalues and the corresponding standardized eigenvectors. After both the FA and PCA loadings are obtained, we compute the closeness measures (to be discussed in the next section); and, at the end of the 100 replications, the average value of the squared canonical correlation (Schneeweiss 1997; Schneeweiss & Mathes 1995) between the FA and PCA loading matrices across the replications is obtained.

4 Closeness

The squared canonical correlation (SCC) between matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}*$ is given by

$$\rho^2\left(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^*\right) = (1/m)\, tr\left\{\left(\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\right)^{-1}\left(\boldsymbol{\Lambda}'\boldsymbol{\Lambda}^*\right)\left(\boldsymbol{\Lambda}^{*\prime}\boldsymbol{\Lambda}^*\right)^{-1}\left(\boldsymbol{\Lambda}^{*\prime}\boldsymbol{\Lambda}\right)\right\}. \tag{1}$$

The Fisher-z transformation of canonical correlation (CC) is given by

$$z = (1/2)\log\left\{\left(1+\rho\left(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^*\right)\right)/\left(1-\rho\left(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^*\right)\right)\right\}, \tag{2}$$

where $\rho(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^*) = \sqrt{\rho^2(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^*)}$. Notice that at the population level for each condition these two measures do not change across replications. We denote this by $z(\hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Lambda}}^*)$ when Eq. (2) is applied to the average sample CC. Clearly, $z(\hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Lambda}}^*)$ measures the closeness between FA and PCA loading estimates. Similarly, we use $z(\hat{\boldsymbol{\Lambda}}, \boldsymbol{\Lambda})$ to measure the closeness between the estimates of FA loadings and the population FA loadings via the average SCC; and use $z(\hat{\boldsymbol{\Lambda}}^*, \boldsymbol{\Lambda})$ to measure the closeness between the estimates of the PCA loadings and the population FA loadings via the average SCC. We employ the Fisher-z transformation because it is good at distinguishing values of SCC that are very close to 1, and also, our prior study (Liang et al. 2015) found a linear relationship between $z(\hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Lambda}}^*)$ and $\log(p)$.

5 Results

We present our results in five three-dimensional Figures (Figs. 1 through 5). Each figure is presented from three different directions to make the results easily visualized. Also, all the figures are given in terms of the Fisher-z transformed average CC on the vertical axis, which is plotted as a function of log(p) and log(N) represented by the two horizontal axes.

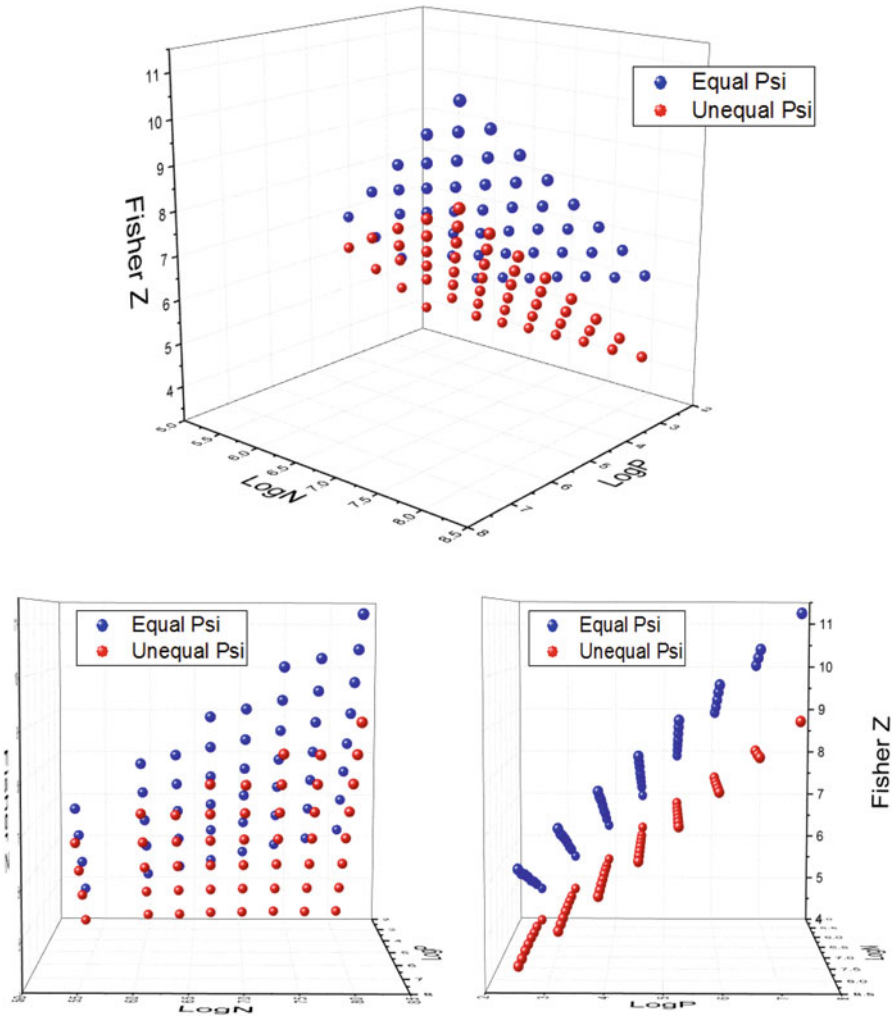


Fig. 1 Fisher-z transformed average canonical correlation between estimated FA and estimated PCA loadings ($\rho(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Lambda}}^*)$) as functions of log($p$) and log($N$): Comparing Equal Psi (in *blue*) and Unequal Psi (in *red*) covariance structures, seen from three difference directions

Figure 1 shows the Fisher-z transformed average CC between the estimated FA and the estimated PCA loadings ($z(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Lambda}}^*)$) as functions of $\log(p)$ and $\log(N)$, where the results under the Equal Psi covariance structure are in blue, and those from the Unequal Psi covariance structure are in red. As Fig. 1 shows, the $z(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Lambda}}^*)$ are always larger under the Equal Psi than under the Unequal Psi covariance structure regardless of the values of $\log(p)$ and $\log(N)$. Also, the $z(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Lambda}}^*)$ under both the Equal Psi and the Unequal Psi covariance structures increase as $\log(p)$ increases.

Second, Fig. 2 describes the Fisher-z transformed average CC between the estimated and the population FA loadings ($z(\hat{\mathbf{\Lambda}}, \mathbf{\Lambda})$) as functions of $\log(p)$ and $\log(N)$. As Fig. 2 shows, the $z(\hat{\mathbf{\Lambda}}, \mathbf{\Lambda})$ are only slightly larger under the Equal Psi than under the Unequal Psi covariance structure. Also, the $z(\hat{\mathbf{\Lambda}}, \mathbf{\Lambda})$ under both the Equal Psi and the Unequal Psi covariance structures clearly increase as $\log(N)$ increases. However, the $z(\hat{\mathbf{\Lambda}}, \mathbf{\Lambda})$ do not change much as $\log(p)$ increases.

Third, Fig. 3 describes Fisher-z transformed average CC between the estimated PCA and the population FA loadings ($z(\hat{\mathbf{\Lambda}}^*, \mathbf{\Lambda})$) as functions of $\log(p)$ and $\log(N)$. As Fig. 3 shows, the $z(\hat{\mathbf{\Lambda}}^*, \mathbf{\Lambda})$ are only slightly larger under the Equal Psi than under the Unequal Psi covariance structure. Also, the $z(\hat{\mathbf{\Lambda}}^*, \mathbf{\Lambda})$ under both the Equal Psi and the Unequal Psi covariance structures clearly increase as $\log(N)$ increases. Furthermore, the $z(\hat{\mathbf{\Lambda}}^*, \mathbf{\Lambda})$ do not change much as $\log(p)$ increases. In short, the results in Fig. 3 are very close to those in Fig. 2.

Now, instead of comparing between the Equal Psi and the Unequal Psi covariance structures, we compare the three Fisher-z transformed average CCs separately for the Equal Psi and the Unequal Psi covariance structures. The quantification of $z(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Lambda}}^*)$, $z(\hat{\mathbf{\Lambda}}, \mathbf{\Lambda})$, and $z(\hat{\mathbf{\Lambda}}^*, \mathbf{\Lambda})$ are expressed in Figs. 4 and 5 as blue, red, and green dots, respectively. First, Fig. 4 describes the three Fisher-z transformed average CCs as functions of $\log(p)$ and $\log(N)$ under the Equal Psi covariance structure. As Fig. 4 shows, the $z(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Lambda}}^*)$ are much larger than the $z(\hat{\mathbf{\Lambda}}, \mathbf{\Lambda})$ and $z(\hat{\mathbf{\Lambda}}^*, \mathbf{\Lambda})$ at every pair of levels of $\log(p)$ and $\log(N)$. Also, $z(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Lambda}}^*)$ increases rapidly as $\log(p)$ increases, but increases more slowly as $\log(N)$ increases. Furthermore, $z(\hat{\mathbf{\Lambda}}^*, \mathbf{\Lambda})$ is only slightly larger than $z(\hat{\mathbf{\Lambda}}, \mathbf{\Lambda})$, and the difference is so small that the $z(\hat{\mathbf{\Lambda}}, \mathbf{\Lambda})$ and $z(\hat{\mathbf{\Lambda}}^*, \mathbf{\Lambda})$ are nearly overlapped. However, as $\log(N)$ increases both the $z(\hat{\mathbf{\Lambda}}, \mathbf{\Lambda})$ and $z(\hat{\mathbf{\Lambda}}^*, \mathbf{\Lambda})$ increase much more slowly than the $z(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Lambda}}^*)$.

Finally, Fig. 5 describes three Fisher-z transformed average CCs as functions of $\log(p)$ and $\log(N)$ under the Unequal Psi covariance structure. The findings are very similar to those from Fig. 4, except for the following two differences. The first difference is that the values of $z(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Lambda}}^*)$ in Fig. 5 are closer to those of $z(\hat{\mathbf{\Lambda}}, \mathbf{\Lambda})$ and $z(\hat{\mathbf{\Lambda}}^*, \mathbf{\Lambda})$ than in Fig. 4. In fact, the $z(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Lambda}}^*)$ almost overlap with $z(\hat{\mathbf{\Lambda}}, \mathbf{\Lambda})$ and $z(\hat{\mathbf{\Lambda}}^*, \mathbf{\Lambda})$ at the smallest values of $\log(p)$ and the largest value of $\log(N)$. The other difference is that when $\log(p)$ is large, the $z(\hat{\mathbf{\Lambda}}, \mathbf{\Lambda})$ almost overlap with $z(\hat{\mathbf{\Lambda}}^*, \mathbf{\Lambda})$, and then the $z(\hat{\mathbf{\Lambda}}, \mathbf{\Lambda})$ become slightly larger than $z(\hat{\mathbf{\Lambda}}^*, \mathbf{\Lambda})$ when $\log(p)$ is small and $\log(N)$ is large.

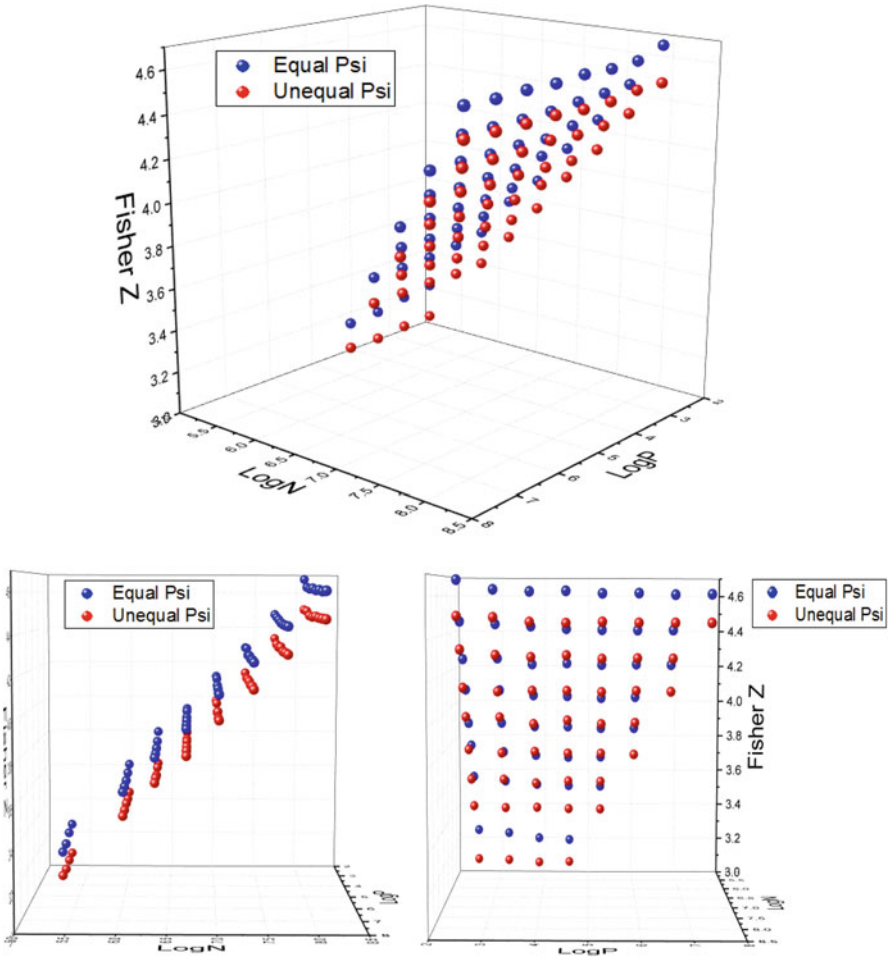


Fig. 2 Fisher-z transformed average canonical correlation between estimated and population FA loadings as functions of log(p) and log(N): Comparing Equal Psi (in *blue*) and Unequal Psi (in *red*) covariance structures, seen from three different directions

6 Discussion

In the first part of the results section (Figs. 1 through 3), we compare the Fisher-z transformed average CC under the Equal Psi condition against that under the Unequal Psi conditions, and find that $z(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Lambda}}^*)$ under the Equal Psi condition is always larger. This is as expected because the Equal Psi covariance structure in the population can be considered as a variant of PCA, also called the FA with equal unique variances (Hayashi & Bentler 2000; Johnstone & Lu 2004, 2009). In fact, there exists an analytical formula to convert $\mathbf{\Lambda}^*$ into $\mathbf{\Lambda}$ for the one factor case with equal unique variances (Bentler & Kano 1990).

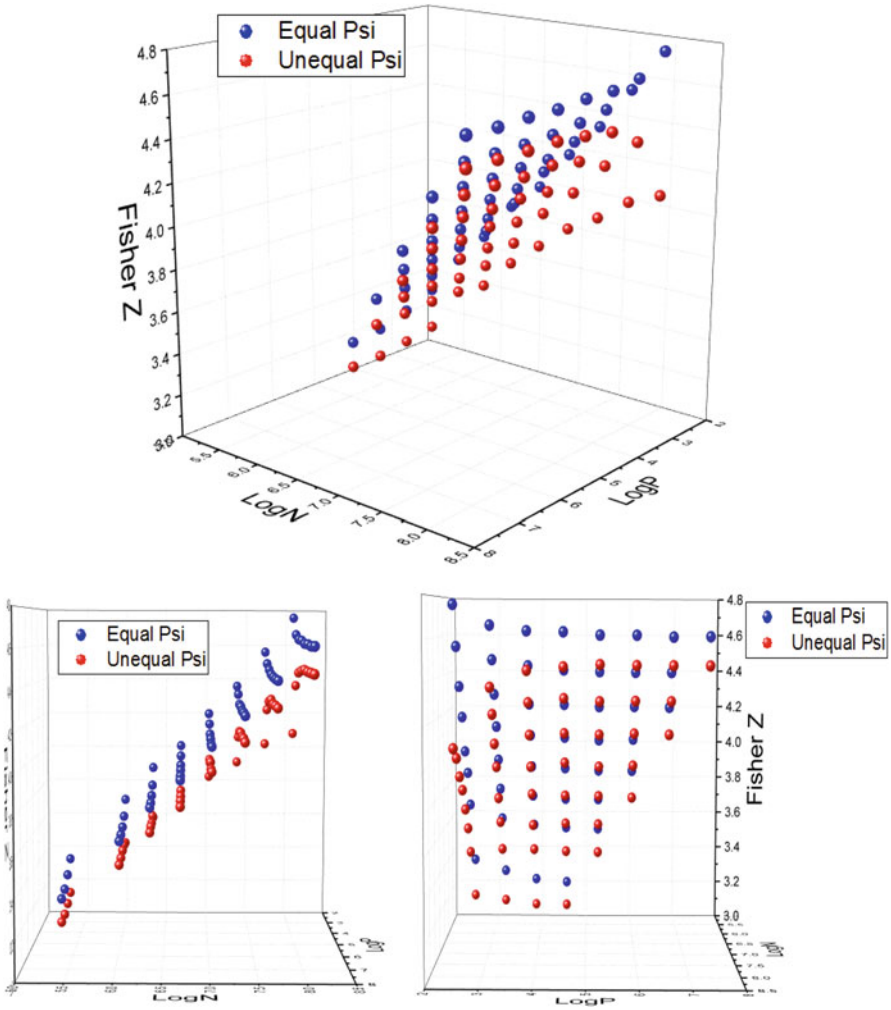


Fig. 3 Fisher-z transformed average canonical correlation between estimated PCA and population FA loadings ($\rho(\hat{\mathbf{\Lambda}}^*, \mathbf{\Lambda})$) as functions of log($p$) and log($N$): Comparing Equal Psi (in *blue*) and Unequal Psi (in *red*) covariance structures, seen from three different directions

It is predicted from the Guttman condition that as p increases, $z(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Lambda}}^*)$ increases. Also, it is predicted from the classical asymptotic theory that as N increases, $z(\hat{\mathbf{\Lambda}}, \mathbf{\Lambda})$ increases. However, what is interesting is the result that as N increases, $z(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Lambda}}^*)$ also slowly increase under the Equal Psi covariance structure. We can probably explain this result as being a consequence of the case with $p/N \to 0$ (Bai & Li 2012; Johnstone & Lu 2004, 2009). More specifically, as log(N) increases with log(p) fixed, we observe the same situation as the case with $p/N \to 0$.

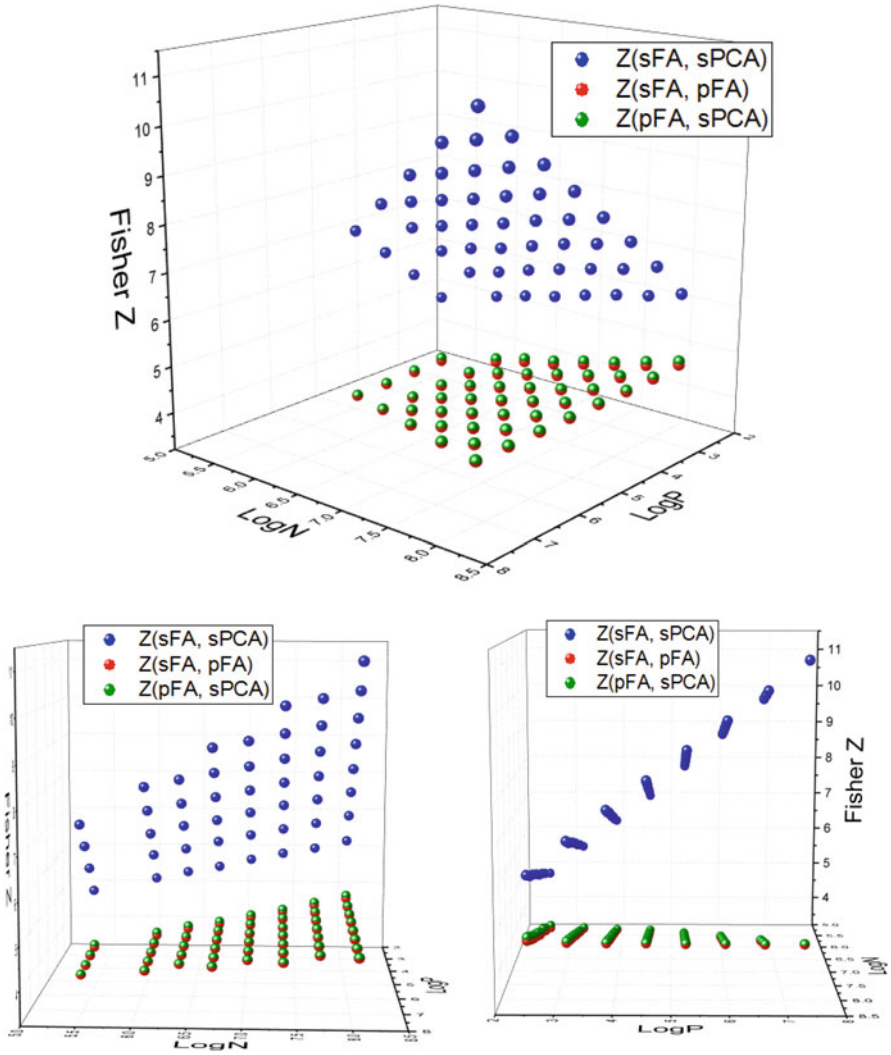


Fig. 4 Three Fisher-z transformed average canonical correlations ((**a**) $z(\hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Lambda}}^*)$ in *blue*, (**b**) $z(\hat{\boldsymbol{\Lambda}}, \boldsymbol{\Lambda})$ in *red*, and (**c**) $z(\hat{\boldsymbol{\Lambda}}^*, \boldsymbol{\Lambda})$ in *green*) as functions of log(p) and log(N) under the Equal Psi covariance structure, seen from three different directions

It is evident that the $z(\hat{\boldsymbol{\Lambda}}^*, \boldsymbol{\Lambda})$ are always slightly larger under the Equal Psi than the Unequal Psi covariance structure. Note that the closeness between $\hat{\boldsymbol{\Lambda}}^*$ and $\boldsymbol{\Lambda}$ can be "decomposed" into the two closeness measures, one between $\hat{\boldsymbol{\Lambda}}^*$ and $\hat{\boldsymbol{\Lambda}}$, and the other between $\hat{\boldsymbol{\Lambda}}$ and $\boldsymbol{\Lambda}$. Because the values of $z(\hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Lambda}}^*)$ are much larger than those of $z(\hat{\boldsymbol{\Lambda}}, \boldsymbol{\Lambda})$, the values of $z(\hat{\boldsymbol{\Lambda}}^*, \boldsymbol{\Lambda})$ are dominated by the smaller values of

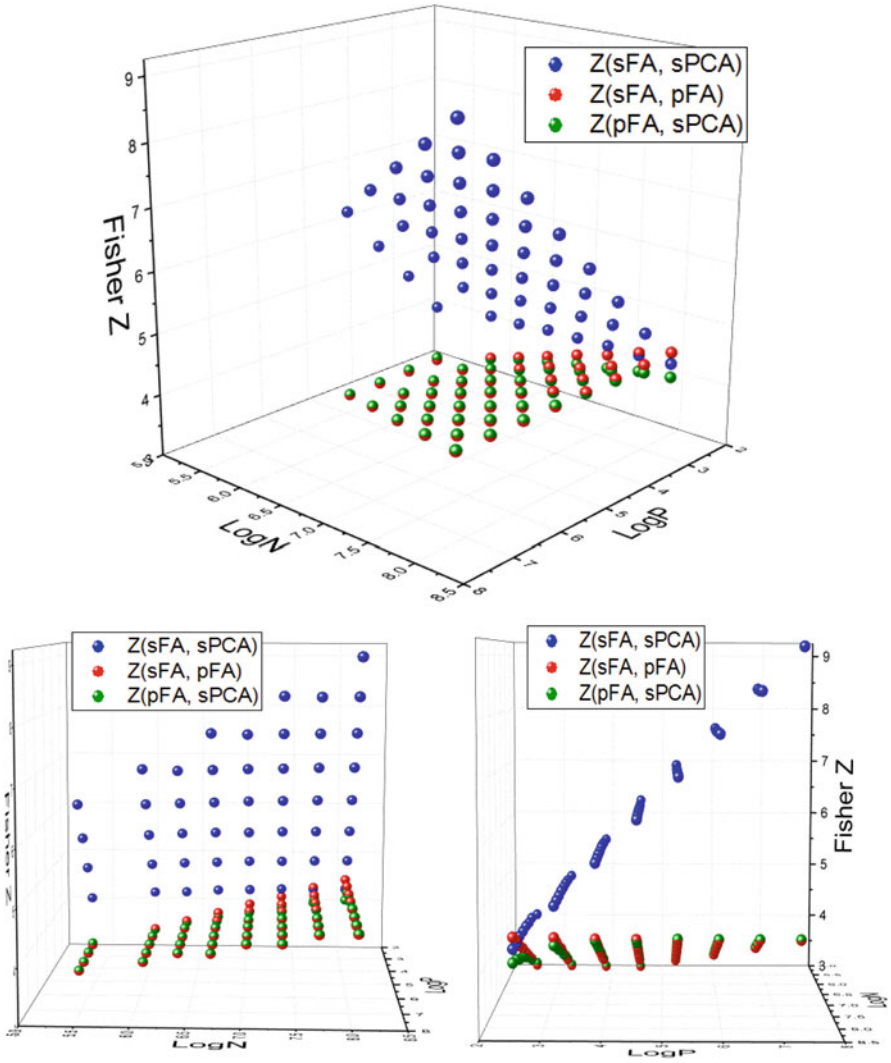


Fig. 5 Three Fisher-z transformed average canonical correlations ((**a**) $z(\hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Lambda}}^*)$ in *blue*, (**b**) $z(\hat{\boldsymbol{\Lambda}}, \boldsymbol{\Lambda})$ in *red*, and (**c**) $z(\hat{\boldsymbol{\Lambda}}^*, \hat{\boldsymbol{\Lambda}})$ in *green*) as functions of log(p) and log(N) under the Unequal Psi covariance structure, seen from three different directions

$z(\hat{\boldsymbol{\Lambda}}, \boldsymbol{\Lambda})$. However, $z(\hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Lambda}}^*)$ still somewhat affects $z(\hat{\boldsymbol{\Lambda}}^*, \boldsymbol{\Lambda})$, and as we have seen, the values of $z(\hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Lambda}}^*)$ are larger under the Equal Psi than under the Unequal Psi covariance structure.

The second half of our results (Figs. 4 and 5) have to do with comparing three Fisher-z's: $z(\hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Lambda}}^*)$, $z(\hat{\boldsymbol{\Lambda}}, \boldsymbol{\Lambda})$, and $z(\hat{\boldsymbol{\Lambda}}^*, \boldsymbol{\Lambda})$, for each covariance structure separately. The results are very close between the Equal Psi and the Unequal Psi

covariance structures. The main finding is that the $z(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Lambda}}^*)$ are larger than $z(\hat{\mathbf{\Lambda}}, \mathbf{\Lambda})$ and $z(\hat{\mathbf{\Lambda}}^*, \mathbf{\Lambda})$ at most combinations of $\log(p)$ and $\log(N)$, with the only exceptions being at the smallest $\log(p)$ under the Unequal Psi covariance structure. The overall results of $z(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Lambda}}^*)$ being larger than $z(\hat{\mathbf{\Lambda}}, \mathbf{\Lambda})$ indicate that the closeness between the estimated FA and the estimated PCA loadings is achieved faster as p increases than is the closeness between the estimated and the population FA loadings as N increases. The implication for the overall results is that, under high-dimensional settings, we should be concerned more about the closeness between the estimated FA loadings and their population counterparts, rather than the closeness between the estimated FA loadings and the estimated PCA loadings.

Although the overall results are very close between the two covariance structures, we also find some interesting differences, that is, $z(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Lambda}}^*)$ are closer in magnitude to $z(\hat{\mathbf{\Lambda}}, \mathbf{\Lambda})$ and $z(\hat{\mathbf{\Lambda}}^*, \mathbf{\Lambda})$ under the Unequal Psi than under the Equal Psi covariance structure, especially when p is small. Again, we suspect that this is related to the fact that the Equal Psi covariance structure is fit perfectly by a variant of PCA.

Obviously, our simulation design is far from being extensive in the sense that the ratios p/N do not include values greater than 0.5, and just two covariance structures were considered. More extensive simulation studies might need to include more varying covariance structures, with different combinations of p and N with p/N being greater than 0.5, in which case we might also need to employ something other than, or in addition to, the ridge approach by Yuan and Chan (2008).

Acknowledgment Ke-Hai Yuan's work was supported by the National Science Foundation under Grant No. SES-1461355. The authors are grateful to Dr. Daniel M. Bolt for his valuable comments on the earlier version of the manuscript.

References

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics, 34*, 122–148.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). New York: Wiley.

Bai, J., & Li, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics, 40*, 436–465.

Beaujean, A. A. (2013). Factor analysis using R. *Practical Assessment, Research & Evaluation, 18*. Retrieved April 12, 2015, from http://pareonline.net/getvn.asp?v=18&n=4.

Bentler, P. M., & Kano, Y. (1990). On the equivalence of factors and components. *Multivariate Behavioral Research, 25*, 67–74.

Guttman, L. (1956). "Best possible" estimates of communalities. *Psychometrika, 21*, 273–286.

Hayashi, K., & Bentler, P. M. (2000). On the relations among regular, equal unique variances, and image factor analysis models. *Psychometrika, 65*, 59–72.

Johnstone, I. M. & Lu, A. Y. (2004). Sparse principal component analysis (Technical report). Department of Statistics, Stanford University.

Johnstone, I. M., & Lu, A. Y. (2009). Consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association, 104*, 682–693.

Krijnen, W. P. (2006). Convergence of estimates of unique variances in factor analysis, based on the inverse sample covariance matrix. *Psychometrika, 71*, 193–199.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method* (2nd ed.). New York: American Elsevier.

Liang, L., Hayashi, K., & Yuan, K.-H. (2015). On closeness between factor analysis and principal component analysis under high-dimensional conditions. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & S.-M. Chow (Eds.) *Quantitative psychology research*: *The 79th Annual Meeting of the Psychometric Society*, *Madison*, *Wisconsin, 2014* (pp. 209–221). New York: Springer.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica, 17*, 1617–1642.

Pourahmadi, M. (2013). *High-dimensional covariance estimation*. New York: Wiley.

Schneeweiss, H. (1997). Factors and principal components in the near spherical case. *Multivariate Behavioral Research, 32*, 375–401.

Schneeweiss, H., & Mathes, H. (1995). Factor analysis and principal components. *Journal of Multivariate Analysis, 55*, 105–124.

Yuan, K.-H. (2013). *Ridge structural equation modeling with large p and/or small N. IMPS2013*. The Netherlands: Arnhem.

Yuan, K.-H., & Chan, W. (2008). Structural equation modeling with near singular covariance matrices. *Computational Statistics and Data Analysis, 52*, 4842–4858.

Remedies for Degeneracy in Candecomp/Parafac

Paolo Giordani and Roberto Rocci

Abstract In many psychological studies variables are measured for some subjects in different conditions. In these cases the available information is stored in a three-way data array. Three-way extensions of Principal Component Analysis have been introduced to summarize such an array through components. Among them a widely used method is the so-called Candecomp/Parafac. Its applicability may be limited by the risk of obtaining degenerate solutions (i.e. solutions with highly collinear, diverging and therefore uninterpretable components). This work focuses on some remedies for avoiding degeneracy based on the use of (hard and/or soft) orthogonality constraints.

Keywords Three-way Data • Candecomp/Parafac • Degeneracy • Orthogonality constraints

1 Introduction

Two-way data usually refer to the scores of a set of subjects with respect to some variables. This information can be stored in a standard two-way matrix, say $\mathbf{X}$ of order $(I \times J)$, where I and J denote the numbers of subjects and variables, respectively. However, in many psychological/behavioural studies it is very common to deal with information regarding some variables collected on a group of subjects during several occasions. We may think about the behaviors of some individuals in different (stressing) situations, the scores of a sample of students with respect to some tests monthly repeated, the levels of several symptoms observed on some patients by different clinicians. In all of the above examples the available information can be stored in a three-way data array, say $\underline{\mathbf{X}}$ of order $(I \times J \times K)$, being K the number of occasions. A three-way data array can be viewed as a collection

P. Giordani (✉)
Sapienza University of Rome, Roma, Italy
e-mail: paolo.giordani@uniroma1.it

R. Rocci
University of Rome "Tor Vergata", Roma, Italy
e-mail: roberto.rocci@uniroma2.it

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer
Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_16

of standard matrices, one for every occasion. See, for more details on three-way analysis, Kroonenberg (2008).

In order to summarize the (preprocessed by centering and normalizing) two-way matrix $\mathbf{X}$ through a limited number of components, say S, Principal Component Analysis (PCA) is the most common tool. We have

$$\mathbf{X} = \mathbf{A}\mathbf{B}' + \mathbf{E}, \tag{1}$$

where $\mathbf{A}$ of order $(I \times S)$ and $\mathbf{B}$ of order $(J \times S)$ are the component matrices for the subjects (scores) and for the variables (loadings), respectively. The term $\mathbf{A}\mathbf{B}'$ provides the best approximation of rank S of $\mathbf{X}$ in the least squares sense. Thus, the optimal component matrices $\mathbf{A}$ and $\mathbf{B}$ are found by minimizing the squared Frobenius norm of the error term $\mathbf{E}$ of order $(I \times J)$:

$$||\mathbf{E}||^2 = ||\mathbf{X} - \mathbf{A}\mathbf{B}'||^2, \tag{2}$$

with respect to $\mathbf{A}$ and $\mathbf{B}$. It is important to note that, without loss of fit, the matrix $\mathbf{A}$ can be chosen, and usually it is, columnwise orthonormal.

Although it is still possible to apply PCA on $\underline{\mathbf{X}}$, for instance by juxtaposing the matrices for every occasion next to each other or averaging the scores across the occasions, this strategy is usually not recommended. This is so because the three-way nature of the data is not taken into account. Hence, the results are almost always incomplete or even inappropriate. For this purpose, several extensions of PCA for three-way data have been proposed in the literature. Among them, the more relevant techniques are the so-called Tucker3 (Tucker 1966) and Candecomp/Parafac (Carroll & Chang, 1970; Harshman 1970), in short T3 and CP, respectively. Differently from the T3, the solution of the CP is unique under mild conditions (Kruskal 1977). This makes the CP widely used in the psychometric domain. Unfortunately, the use of CP may be prevented due to the risk of obtaining degenerate solutions (i.e. solutions with highly collinear, diverging and therefore uninterpretable components).

The current work deals with CP degeneracy. In the next sections, after recalling the CP model and defining the concept of degeneracy with its undesirable effects, we discuss some strategies to avoid degeneracy based on the use of (hard and/or soft) orthogonality constraints. Their use in practice is then shown by means of the application to real data. Some considerations and remarks conclude the paper.

2 Candecomp/Parafac

The Candecomp/Parafac (CP) model has been independently proposed by Carroll and Chang (1970) and Harshman (1970) for summarizing $\underline{\mathbf{X}}$ by extracting a limited number of components, say S. The CP model is a three-way extension of PCA where a new component matrix holding the scores of the occasions with respect to the

components is added. Such a matrix of order $(K \times S)$ is denoted by $\mathbf{C}$. Starting from (1) the CP model can be expressed as

$$\mathbf{X}_A = \mathbf{A}(\mathbf{C} \odot \mathbf{B})' + \mathbf{E}_A, \tag{3}$$

where $\mathbf{X}_A$ is the matrix of order $(I \times JK)$ obtained juxtaposing next to each other the frontal slabs of $\underline{\mathbf{X}}$, i.e. the data matrices corresponding to the different occasions. The matrix $\mathbf{E}_A$ denotes the error term and the symbol $\odot$ is the Khatri-Rao product between two matrices ($\mathbf{C} \odot \mathbf{B} = [\mathbf{c}_1 \otimes \mathbf{b}_1, \ldots, \mathbf{c}_S \otimes \mathbf{b}_S]$ where $\otimes$ is the Kronecker product of matrices).

CP can be seen as a constrained version of the more general Tucker3 (T3) model (Tucker 1966). T3 can be formalized as

$$\mathbf{X}_A = \mathbf{A}\mathbf{G}_A(\mathbf{C} \otimes \mathbf{B})' + \mathbf{E}_A, \tag{4}$$

where the component matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ have order $(I \times P)$, $(J \times Q)$ and $(K \times R)$, respectively, being P, Q and R the numbers of components for the subjects, the variables and the occasions, respectively. Hence, differently from CP, different numbers of components can be chosen for the three modes (subjects, variables or occasions). $\mathbf{G}_A$ is the matrix of order $(P \times QR)$ obtained juxtaposing next to each other the frontal slabs of the so-called core array $\underline{\mathbf{G}}$ of order $(P \times Q \times R)$. The elements of $\underline{\mathbf{G}}$, denoted by g_{pqr}, $p = 1, \ldots, P$, $q = 1, \ldots, Q$, $r = 1, \ldots, R$, provide the information about the triple interactions among the components of the three modes. The CP model with S components can be obtained setting $P = Q = R = S$ and $g_{pqr} = 1$ when $p = q = r = s$ and 0 otherwise, i.e., $\underline{\mathbf{G}} = \underline{\mathbf{I}}$, being $\underline{\mathbf{I}}$ the three-way identity array. Let $\mathbf{I}_A$ denote the matrix obtained juxtaposing next to each other the frontal slabs of $\underline{\mathbf{I}}$. It is easy to see that $\mathbf{I}_A(\mathbf{C} \otimes \mathbf{B})' = (\mathbf{C} \odot \mathbf{B})'$ and, therefore, when $\underline{\mathbf{G}} = \underline{\mathbf{I}}$, (3) and (4) coincide. In practice, this implies that in CP only a subset of triple interactions among components are allowed, i.e., when $p = q = r = s$.

In CP the data are usually preprocessed. In three-way analysis the preprocessing step is more complex than the one in the standard two-way case. In fact, the data can still be preprocessed by centering and normalizing to eliminate artificial differences in level and scale. Nonetheless, there exist several alternatives to do it. One should decide which mode(s) should be considered for centering and/or normalizing. Different choices of preprocessing affect the obtained results. See, for further details, Bro and Smilde (2003) and Harshman and Lundy (1984).

The CP solution can be found by minimizing the following loss function

$$||\mathbf{E}_A||^2 = ||\mathbf{X}_A - \mathbf{A}(\mathbf{C} \odot \mathbf{B})'||^2, \tag{5}$$

with respect to the three component matrices. Several Alternating Least Squares (ALS) algorithms have been proposed in the literature. The most common one consists of solving three multivariate linear regression problems with respect to each component matrix iteratively upon convergence. For a comparative assessment of the various ALS algorithms refer to Tomasi and Bro (2006).

Some of these algorithms are implemented in various software. We mention the program `3WayPack` (Kroonenberg 1996), in `Matlab` the *N*-`way` toolbox (Andersson & Bro, 2000), the `Tensor` toolbox (Bader et al. 2015) and the `Three-way` `m`-files (see, e.g., Kiers & Van Mechelen, 2001), in `R` the package `ThreeWay` (Giordani, Kiers, & Del Ferraro, 2014).

The goodness of fit of the CP model is expressed in terms of the residual sum of squares

$$\left(1 - \frac{||\mathbf{X}_A - \mathbf{A}(\mathbf{C} \odot \mathbf{B})'||^2}{||\mathbf{X}_A||^2}\right) 100. \tag{6}$$

Values of (6) close to 100 mean that CP fits the data very well.

The CP model is widely used for its valuable properties. Among them we mention the uniqueness of its solution up to scaling and permuting the columns of the component matrices (Kruskal 1977). In particular, the uniqueness property holds if

$$2S + 2 \leq k\text{−rank}(\mathbf{A}) + k\text{−rank}(\mathbf{B}) + k\text{−rank}(\mathbf{C}), \tag{7}$$

where the k−rank of a matrix is the largest number k such that the columns are linearly independent in every subset of k columns. This condition has been refined by several authors (Domanov & De Lathauver, 2013a; 2013b; Jiang & Sidiropoulos, 2004; Stegeman 2009a; Stegeman, ten Berge, & De Lathauwer, 2006). The uniqueness property does not hold for T3.

Another attractive property of CP is that the solution with S components provides the best approximation of tensorial rank S of $\underline{\mathbf{X}}$ (Kruskal 1977), similarly to the PCA solution with S components giving the best approximation of (matrix) rank S of $\mathbf{X}$.

2.1 Degeneracy

Despite its valuable properties, the applicability of CP may be prevented by the risk of degenerate solutions. In the literature, this phenomenon has been firstly reported by Harshman and Lundy (1984). A CP solution is affected by degeneracy when (some of) the extracted components are diverging and the columns of the component matrices are highly collinear. Despite this, the sum of the components contribute to improve the fit of the CP model. Therefore, although the fit of a degenerate solution can be very good, the extracted components are uninterpretable and, thus, have no practical use. Another common feature of CP degeneracy is the abnormal computation time needed to minimize the loss function in (5). To assess whether a solution is degenerate the minimum triple cosine between pairs of components is computed (see, e.g., Kroonenberg 2008). The triple cosine between components s and s' is defined as

$$\begin{aligned} \text{tcos}(s, s') &= \cos(\mathbf{a}_s \otimes \mathbf{b}_s \otimes \mathbf{c}_s, \mathbf{a}_{s'} \otimes \mathbf{b}_{s'} \otimes \mathbf{c}_{s'}) = \frac{(\mathbf{a}_s \otimes \mathbf{b}_s \otimes \mathbf{c}_s)'(\mathbf{a}_{s'} \otimes \mathbf{b}_{s'} \otimes \mathbf{c}_{s'})}{||\mathbf{a}_s \otimes \mathbf{b}_s \otimes \mathbf{c}_s|| ||\mathbf{a}_{s'} \otimes \mathbf{b}_{s'} \otimes \mathbf{c}_{s'}||} \\ &= \cos(\mathbf{a}_s, \mathbf{a}_{s'})\cos(\mathbf{b}_s, \mathbf{b}_{s'})\cos(\mathbf{c}_s, \mathbf{c}_{s'}) = \frac{\mathbf{a}_s'\mathbf{a}_{s'}}{||\mathbf{a}_s|| ||\mathbf{a}_{s'}||} \frac{\mathbf{b}_s'\mathbf{b}_{s'}}{||\mathbf{b}_s|| ||\mathbf{b}_{s'}||} \frac{\mathbf{c}_s'\mathbf{c}_{s'}}{||\mathbf{c}_s|| ||\mathbf{c}_{s'}||}. \end{aligned} \tag{8}$$

If the minimum triple cosine is lower than -0.99, it is usually assumed that the solution suffers from degeneracy.

For the sake of completeness, note that two types of degeneracy have been defined in the literature. The above described one (usually referred to as Type I) is the most problematic one. Type II degeneracy (Mitchell & Burdick, 1994) deals with the case in which the CP algorithm is extremely slow because the current CP solution is close to be degenerate, but then emerges and converges to a non-degenerate solution.

The reason for the occurrence of (Type I) degeneracy has been deeply studied by several authors. See, for instance, De Silva and Lim (2008), Giordani and Rocci (2013a), Giordani and Rocci (2013b), Giordani and Rocci (2016), Harshman and Lundy (1984), Krijnen, Dijkstra, and Stegeman (2008), Kruskal, Harshman, and Lundy (1989), Lim and Comon (2009), Lundy, Harshman, and Kruskal (1989), Paatero (2000), Rocci and Giordani (2010), Stegeman (2006), Stegeman (2007), Stegeman (2008), Stegeman (2009b), Stegeman (2012), Stegeman (2013), Stegeman (2014), Stegeman and De Lathauwer (2009), ten Berge, Kiers, and De Leeuw (1988). It is nowadays recognized that it depends on the loss function in (5) that may not have the minimum, but an infimum. A formal proof of the phenomenon of CP degeneracy due to the non-existence of an optimal CP solution can be found in Krijnen, Dijkstra, and Stegeman (2008). In terms of the concept of rank, differently from the two-way case, we have that the best approximation of $\underline{\mathbf{X}}$ of tensorial rank S does not always exist (an example can be seen in, for instance, Rocci & Giordani, 2010). This is so because there exist arrays of tensorial rank $T(> S)$ that are the limit of sequences of arrays of tensorial rank S. If S is the minimal integer such that this holds, then the arrays have tensorial rank T, but border rank S (Bini 1980).

In order to solve the CP degeneracy problem various remedies have been proposed. Generally speaking, they consist of reconsidering the CP minimization problem in such a way to guarantee the existence of the minimum. The most common strategy to achieve this goal is to add suitable constraints on the component matrices.

3 Remedies Against Degeneracy

Two types of constraints are usually imposed to avoid degeneracy at the cost of losing a certain amount of fit. One can either impose the non-negativity of all the component matrices (see, e.g., Lim & Comon, 2009) or the orthogonality of one of the component matrices (see, e.g., Harshman & Lundy, 1984). The latter type of constraints, being standard in PCA, is more common in the psychometric field and, therefore, will be investigated in this work. The former one is rather unusual because

it requires the non-negativity of the data array with a meaningful and non-arbitrary zero point.

3.1 Candecomp/Parafac with Orthogonality Constraints

The Candecomp/Parafac with orthogonality constraints (CP-Orth) (Harshman & Lundy, 1984) can be written as

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{B},\mathbf{C}} ||\mathbf{X}_A - \mathbf{A}(\mathbf{C} \odot \mathbf{B})'||^2, \\ &\text{s.t.} \quad \mathbf{A}'\mathbf{A} = \mathbf{I}, \end{aligned} \tag{9}$$

being $\mathbf{I}$ the identity matrix. In (9), without loss of generality, we assume that the orthogonality constraints refer to the component matrix $\mathbf{A}$, which is also required to be columnwise normalized, i.e., $\mathbf{A}$ is columnwise orthonormal. The constraints guarantee that the problem in (9) always has the minimum. A formal proof can be found in Krijnen, Dijkstra, and Stegeman (2008).

Although CP-Orth could be seen as the final solution for solving the CP degeneracy problem, it may fail from a practical point of view. In fact, one may doubt whether it is reasonable to interpret the presence of degeneracy as an indication of the presence of orthogonal components underlying the data. The orthogonality constraints are arbitrarily imposed without a formal assessment about its soundness for the data at hand.

3.2 Candecomp/Parafac with Lasso Constraints

For the above-mentioned reason, the so-called Candecomp/Parafac with Lasso constraints (CP-Lasso) has been proposed (Giordani & Rocci, 2013a). It avoids CP degeneracy by softening the orthogonality constraints as follows:

$$\begin{aligned} &\min_{\mathbf{Q},\mathbf{R},\mathbf{B},\mathbf{C}} \left\|\mathbf{X}_A - \mathbf{Q}\mathbf{R}\,(\mathbf{C} \odot \mathbf{B})'\right\|^2, \\ &\text{s.t.} \quad \mathbf{R} \text{ upper triangular with } \operatorname{diag}(\mathbf{R}) = \mathbf{1}, \\ &\qquad \sum_{s=1}^{S} \sum_{s'=s+1}^{S} |r_{ss'}| \leq \tau,\ \mathbf{Q}'\mathbf{Q} = \mathbf{I}. \end{aligned} \tag{10}$$

The CP-Lasso is based on the QR-factorization of $\mathbf{A}$,

$$\mathbf{A} = \mathbf{Q}\mathbf{R}, \tag{11}$$

where $\mathbf{Q}$ is columnwise orthonormal and $\mathbf{R}$ is an upper triangular matrix with 1's in its main diagonal (this can always be done by using the scaling indeterminacy of the column norms of the component matrices). The peculiarity of CP-Lasso is to impose that the sum of the upper and off-diagonal elements of $\mathbf{R}$ in absolute value must be lower than a pre-specified quantity denoted by $\tau(\geq 0)$. This constraint takes inspiration from the well-known Lasso (acronym of Least Absolute Shrinkage and Selection Operator) procedure widely used in regression (Tibshirani 1996). Let $\mathbf{y}$ and $\mathbf{X}$ denote the vector of the response and the matrix of the J explanatory variables, respectively. The Lasso regression can be written as

$$\begin{aligned} &\min_{\mathbf{b}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2, \\ &\text{s.t. } \sum_{j=1}^{J} |b_j| \leq \tau, \end{aligned} \tag{12}$$

being $\mathbf{b}$ the vector of the regression coefficients. A nice property of Lasso is the tendency to produce some estimated regression coefficients equal to 0. The problem in (12) can be solved iteratively.

Any finite value of τ guarantees that the CP-Lasso solution is not affected by degeneracy. If $\tau = 0$, then $\mathbf{R} = \mathbf{I}$ and, therefore, $\mathbf{A} = \mathbf{Q}$, i.e. CP-Lasso coincides with CP-Orth. An ALS algorithm can be used for minimizing (10). The update of $\mathbf{R}$ boils down to solve a particular Lasso regression problem as the one recalled in (12). This implies that in CP-Lasso it is frequent to estimate some upper and off-diagonal elements of $\mathbf{R}$ by 0. It can be shown that if the generic element $r_{ss'} = 0$ $(s < s')$, then components s and s' are orthogonal or partially orthogonal given the first $s - 1$ (the residuals of the projections of components s and s' on the subspace spanned by the first $s - 1$ components are orthogonal). In general a low value of $|r_{ss'}|$ means that the involved components are nearly (partially) orthogonal. In other words, CP-Lasso softens the orthogonality constraints by stimulating the components to be characterized by a low level of (partially) orthogonality and, for the Lasso geometry, to be pairwise (partially) orthogonal. See, for more details, Giordani and Rocci (2013a).

3.3 Candecomp/Parafac with Ridge Regularization

A side-effect of CP-Lasso is its computational complexity because, as we pointed out, the update of $\mathbf{R}$ requires an iterative solution. Of course, this may increase the computation time of the ALS algorithm for minimizing (10). This limitation can be overcome by replacing the Lasso constraints on $\mathbf{R}$ with a ridge regularization term.

In regression, the standard ridge problem can be formulated as

$$\min_{\mathbf{b}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 + \lambda \|\mathbf{b}\|^2, \tag{13}$$

with $\lambda \geq 0$. As is well-known, the solution to the problem in (13) is given by

$$\mathbf{b} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}'\mathbf{y}. \tag{14}$$

The use of a ridge regularization in the CP domain leads to the Candecomp/Parafac with ridge regularization (CP-Ridge) proposed by Giordani and Rocci (2013b). It can be expressed as

$$\begin{aligned} &\min_{\mathbf{Q},\mathbf{R},\mathbf{B},\mathbf{C}} \left\| \mathbf{X}_A - \mathbf{Q}\mathbf{R}\left(\mathbf{C} \odot \mathbf{B}\right)' \right\|^2 + \lambda \left\| \mathbf{R} - \mathbf{I} \right\|^2, \\ &\text{s.t.} \quad \mathbf{R} \text{ upper triangular with } \text{diag}(\mathbf{R}) = \mathbf{1},\ \mathbf{Q}'\mathbf{Q} = \mathbf{I}. \end{aligned} \tag{15}$$

We can see that the Lasso-based constraints on $\mathbf{R}$ in (10) vanish, while a ridge regularization term involving $\mathbf{R}$ appears. The role of the regularization term is to limit as much as possible the sum of squares of the upper and off-diagonal elements of $\mathbf{R}$. In this case, the tuning of such a new term is played by the non-negative parameter λ. Any positive value of λ avoids the occurrence of degeneracy. The lower the value of λ, the softer the orthogonality constraints are, until $\lambda = 0$, where CP-Ridge reduces to the unconstrained CP problem.

An ALS algorithm can be used for obtaining the minimum of (15). This algorithm is more efficient than the one for CP-Lasso because the update of $\mathbf{R}$ consists of solving a particular ridge regression problem, the closed form solution of which is recalled in (14). The updates of the remaining matrices do not vary with respect to CP-Lasso.

Differently from the Lasso, the ridge regularization is only a shrinking procedure and, hence, the selection property of Lasso does no longer hold. This implies that it is very difficult to have pairwise (partially) orthogonal components in CP-Ridge. For practical purposes, the CP-Lasso is recommended if the researcher seeks for (partially) orthogonal components, otherwise the more computationally efficient CP-Ridge method should be considered.

3.4 Candecomp/Parafac with Singular Value Decomposition Penalization

CP-Lasso and CP-Ridge are based on the QR-factorization of $\mathbf{A}$ and, in particular, act on the upper and off-diagonal elements of $\mathbf{R}$. In this way, they prevent the condition number of $\mathbf{A}$, $cn(\mathbf{A})$, defined as the ratio between the maximum and minimum singular values of $\mathbf{A}$, from going to infinity. In fact, taking into account that $\mathbf{Q}$ is columnwise orthonormal, we have

$$cn(\mathbf{A}) = cn(\mathbf{R}). \tag{16}$$

From (16) it follows that shrinking the upper and off-diagonal elements of $\mathbf{R}$ allows us to handle $cn(\mathbf{A})$. This is a very relevant goal taking into account that a CP solution is degenerate when

$$cn(\mathbf{A}) \to \infty. \tag{17}$$

By limiting the upper and off-diagonal elements of $\mathbf{R}$ we implicitly provide an upper bound for $cn(\mathbf{A})$ preventing it from being infinite. It is interesting to note that the minimum value of $cn(\mathbf{A})$ is equal to 1 and it is obtained if and only if $\mathbf{A}$ is columnwise orthonormal, as is in the CP-Orth case. This further clarifies how CP-Orth is generally very strict. The orthogonality constraints on $\mathbf{A}$ aim at avoiding (17). Although any finite value of $cn(\mathbf{A})$ would be acceptable for getting a non-degenerate solution, as is for CP-Lasso and CP-Ridge, CP-Orth avoids (17) by imposing the too strict condition $cn(\mathbf{A}) = 1$.

The previous comments highlight that it crucial to bind $cn(\mathbf{A})$ for solving the CP degeneracy problem. For this reason, it would be very intuitive a method based directly on the singular values of $\mathbf{A}$. The Candecomp/Parafac with Singular Value Decomposition penalization (CP-SVD) fills this gap (Giordani & Rocci, 2016). The starting point is the Singular Value Decomposition (SVD) of $\mathbf{A}$

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}', \tag{18}$$

where $\mathbf{U}$ and $\mathbf{V}$ are the matrices of, respectively, the left and right singular vectors of $\mathbf{A}$ such that $\mathbf{U}'\mathbf{U} = \mathbf{I}$ and $\mathbf{V}'\mathbf{V} = \mathbf{I}$ and $\mathbf{D}$ is the diagonal matrix holding the singular values of $\mathbf{A}$ in the main diagonal. CP-SVD can be formalized as

$$\begin{aligned} \min_{\mathbf{U},\mathbf{D},\mathbf{V},\mathbf{B},\mathbf{C}} & \left\| \mathbf{X}_A - \mathbf{U}\mathbf{D}\mathbf{V}' \left(\mathbf{C} \odot \mathbf{B}\right)' \right\|^2 + \lambda \left\| \mathbf{D} - \mathbf{I} \right\|^2, \\ \text{s.t.} \quad & \mathbf{D} \text{ diagonal}, \mathbf{U}'\mathbf{U} = \mathbf{I}, \mathbf{V}'\mathbf{V} = \mathbf{I}. \end{aligned} \tag{19}$$

By inspecting (19) we can see that the loss function is the sum of two terms. The first one is the standard CP loss, whereas the second one is a penalty term that is equal to zero if and only if $cn(\mathbf{A}) = 1$. The penalty term represents a measure of non-orthonormality of $\mathbf{A}$ because it compares its singular values with those of an orthonormal matrix. In order to tune the relevance of the penalty term the pre-specified non-negative coefficient λ is considered. A suitable choice of λ prevents degeneracy. The problem in (19) can be solved by means of an ALS algorithm. For further details refer to Giordani and Rocci (2016).

4 Application

In this section we applied the CP-SVD method to the so-called TV data (Lundy, Harshman, & Kruskal, 1989). The data are available in the `R` package `ThreeWay` (Giordani, Kiers, & Del Ferraro, 2014). The data array contains the ratings given

by 30 students on a set of 15 TV shows with respect to 16 bipolar scales. It is well-known that the data array admits a degenerate CP solution with $S = 3$ components. The minimum triple cosine approaches to -1 and two extracted components are highly collinear and uninterpretable. This result has been found by several authors. See, for instance, Lundy, Harshman, and Kruskal (1989) and Stegeman (2014).

Before analyzing the data, we preprocessed them by centering across TV shows and normalizing within students. We then run CP-SVD in order to obtain a solution not affected by degeneracy setting λ according to the selection procedure given by Giordani and Rocci (2016). The SVD is computed on the component matrix for the ratings. The fit of CP-SVD, expressed in terms of (6), is slightly lower than that of CP (51.33 % for CP-SVD and 52.26 % for CP, hence -0.93 %). This highlights a very good performance of CP-SVD, in comparison with CP-Orth, the fit of which is 48.38 % (-2.95 % with respect to the CP-SVD one).

These results stimulate us to investigate the CP-SVD solution. First of all, it does not suffer from degeneracy (the minimum triple cosine is -0.27). Since the matrix for the students (not reported here) contains all non-negative scores, the component matrices for the ratings and the TV shows allow us to interpret the extracted components. These two component matrices are reported in Tables 1 and 2.

Component 1 can be interpreted as 'Sob stories' since it is mainly related to The Waltons, Little House on the Prairie and, with negative sign, to Saturday Night Live and Mash. Such a component is related to TV shows mainly recognized as Uninteresting, Boring, Intellectually dull, Uninformative and Not funny. Component 2 is dominated by Football and is therefore labeled as 'Football (vs others)'. Football is considered to be Callous, Insensitive, Shallow, Crude, Violent, in contrast to the

Table 1 Component matrix for the ratings (scores higher than 0.30 in absolute value are in bold face)

Ratings	Component 1	Component 2	Component 3
Thrilling-Boring	**0.34**	0.07	0.14
Intelligent-Idiotic	**0.31**	0.20	**0.36**
Erotic-Not erotic	0.11	0.02	−0.20
Sensitive-Insensitive	0.10	**0.39**	0.15
Interesting-Uninteresting	**0.38**	0.20	0.24
Fast-Slow	0.28	0.00	0.09
Intellectually stimulating-Intellectually dull	**0.33**	0.22	**0.34**
Violent-Peaceful	0.11	**−0.30**	0.09
Caring-Callous	0.08	**0.39**	0.14
Satirical-Not satirical	0.28	0.18	−0.26
Informative-Uninformative	**0.31**	0.18	**0.37**
Touching-Leave me cold	0.19	**0.38**	0.12
Deep-Shallow	0.23	**0.31**	0.28
Tasteful-Crude	0.15	**0.30**	0.26
Real-Fantasy	0.21	0.08	**0.42**
Funny-Not funny	**0.30**	0.28	−0.21

Note Negative scores refer to the left side of the bipolar scale

Table 2 Component matrix for the TV shows (scores higher than 0.30 in absolute value are in bold face)

Ratings	Component 1	Component 2	Component 3
Mash	**−0.32**	−0.22	0.14
Charlies Angels	0.16	0.21	0.18
All in the Family	−0.13	−0.14	0.22
60 min	−0.16	−0.01	**−0.31**
The Tonight Show	−0.27	−0.03	0.21
Let's Make a Deal	0.23	0.20	0.17
The Waltons	**0.51**	**−0.41**	−0.11
Saturday Night Live	**−0.35**	0.29	**0.40**
News	−0.10	0.22	**−0.36**
Kojak	−0.01	0.23	0.04
Mork and Mindy	−0.05	−0.25	**0.36**
Jacques Cousteau	0.06	−0.12	**−0.41**
Football	−0.20	**0.51**	−0.13
Little House on the Prairie	**0.51**	**−0.39**	−0.08
Wild Kingdom	0.14	−0.08	**−0.32**

other TV shows, in particular The Waltons and Little House on the Prairie having the lowest component scores. Finally, Component 3 is positively related to Saturday Night Live and Mork and Mindy and negatively related to Jacques Cousteau, News, Wild Kingdom and 60 min. The TV shows with the highest component scores are described as Fantasy, Uninformative and Idiotic. Therefore, this component seems to reflect the duality between 'Frivolous vs Factual' TV shows.

The above-described CP-SVD solution resembles to some extent those obtained by Lundy, Harshman, and Kruskal (1989) and Stegeman (2014) although the three solutions are not fully comparable. This depends on the different preprocessing steps adopted by the authors. In fact, Stegeman (2014) centers across TV shows and ratings and normalizes within students, while no details about preprocessing are reported in Lundy, Harshman, and Kruskal (1989). They firstly analyze the TV data by CP-Orth (imposing orthogonality constraints on the component matrix for the ratings) with $S = 3$ components and then estimate the corresponding T3 core by solving an ordinary regression problem in order to discover possible interactions among the components. The strategy, called PFCORE, consists in extracting the orthogonal CP components and then computing (in a single step) the core array. PFCORE is motivated by the assumption that the T3 structure in the data may cause degenerate CP solution (Kruskal, Harshman, & Lundy 1989). By means of PFCORE, the data are expressed in terms of a T3-based model, more general than CP.

The obtained components are interpreted as 'Humor', 'Sensitivity' and 'Violence'. The component labeled 'Humor' is mainly related to Mork and Mindy, Saturday Night Live, Charlie's Angels, Let's Make a Deal (positive scores) and to Jacques Cousteau, News, 60 Minutes, The Waltons (negative scores). The TV

shows with positive scores are rated by the students highly Satirical, Funny, Erotic, Uninformative, Intellectually dull, Idiotic, Fantasy, Shallow and Violent. The opposite comment holds for the TV shows with negative scores. The interpretation of the second component depends on the high scores of Caring, Sensitive, Touching, Boring, Slow and Peaceful. These ratings well characterize TV shows such as The Waltons, Little House on the Prairie (positive scores) and Football, News, Saturday Night Live (negative scores). Finally, the third component is interpreted as 'Violence' because Violent and, to a lesser extent, Not Funny, Fast and Real have high component scores together with TV shows like Football, Charlie's Angels and Kojak. In contrast, negative component scores pertain to Mork and Mindy, The Waltons, Little House on the Prairie and All in the Family.

The analysis of the PFCORE core highlights several interactions among components that cannot be discovered by CP. Since such interactions involve the component labeled 'Humor', Kruskal, Harshman, and Lundy (1989) argue that these are related to differences in the students' sense of humor. These differences cannot be discovered by CP and, hence, the need for T3-based models arises.

Stegeman (2014) analyzes the TV data by means of the so-called CP-Limit method (Stegeman 2012; Stegeman, 2013). The idea underlying CP-Limit is based on the evidence that the best approximation of rank S of an array in the least squares sense belongs to a boundary point of the set of arrays of rank S. If this boundary point has rank at most S, then the optimal CP solution with S components is found; otherwise, degeneracy occurs. This is so because the CP algorithm aims at reaching a boundary point having rank larger than S. Of course, this limit point cannot be hit because the rank of the CP solution can be at most S. For all of these reasons, the CP-Limit enlarges the set of the feasible solutions admitting boundary limit points with rank larger than S. The resulting CP-Limit solution is no longer a CP decomposition and, as in Rocci and Giordani (2010) for $S = 2$, it is represented as a T3 decomposition with a constrained core, where some pre-specified core elements are zero. The location of the zero-constrained elements does not depend on S, but on the number of groups of CP diverging components and on the number of diverging components in each group. For further details on the CP-Limit method refer to Stegeman (2012) and Stegeman (2013).

By applying CP-Limit to the (preprocessed) TV data, Stegeman (2014) find three components. Although the obtained component matrices are not reported, these are interpreted as 'Humor', 'Sensitive' and 'Violence' consistently with Lundy, Harshman, and Kruskal (1989). Even if the extracted components are interpreted in the same way, the CP-Limit components are not constrained to be orthogonal. Another difference between the two solutions is that the core elements of CP-Limit and PF-CORE noticeably disagree denoting different kinds of interactions among components.

All in all, we can thus state that each of the three methods discovers a specific "picture" of the TV data. However, the three solutions are consistent to some extent. In fact, the three components extracted by using CP-Limit and PF-CORE (CP-Orth) can be interpreted in the same way. The components obtained by means of CP-SVD are labeled in a different way. Nonetheless, by observing the scales and the TV shows playing a more relevant role in the component interpretations, some relationships are clearly visible. Specifically, the CP-SVD components interpreted

as 'Sob stories', 'Football (vs others)' and 'Frivolous vs Factual' appears to be closely related to the PF-CORE components labeled 'Sensitive', 'Violence' and 'Humor', respectively.

5 Final Remarks

In this paper we have discussed some tools for solving the CP degeneracy problem. The intuition behind all these methods is to add hard or soft orthogonality constraints to the CP minimization problem in order to guarantee the existence of the optimal solution. Although all of these strategies work well from a mathematical point of view, in practice we recommend to adopt remedies such as CP-Lasso, CP-Ridge or CP-SVD, where the constraints are suitably softened. In this respect, another possibility is given by CP-Limit where the CP degeneracy problem is solved by enlarging the set of feasible solutions (the set of arrays with rank at most S). This is done by admitting boundary points of the set having rank larger than S. It allows us to highlight the existing differences between CP-Lasso, CP-Ridge and CP-SVD on the one side and CP-Limit on the other side. CP-Limit looks for a T3 decomposition with several zero core elements. Therefore, the obtained solution is no longer a CP solution in a strict sense because it has rank larger than S. Conversely, the CP-Lasso, CP-Ridge or CP-SVD solutions are particular CP solutions of rank S not suffering from degeneracy thanks to the corresponding regularization terms.

Acknowledgements The first author gratefully acknowledges the grant FIRB2012 entitled "Mixture and latent variable models for causal inference and analysis of socio-economic data" for the financial support.

References

Andersson, C. A., & Bro, R. (2000). The *N*-way Toolbox for MATLAB. *Chemometrics and Intelligent Laboratory Systems, 52*, 1–4. http://www.models.life.ku.dk/nwaytoolbox. Cited 29 Jan 2016.

Bader, B. W., Kolda, T. G., Sun, J., et al. (2015). MATLAB Tensor Toolbox Version 2.6. http://www.sandia.gov/~tgkolda/TensorToolbox/. Cited January 29, 2016.

Bini, D. (1980). Border rank of a $p \times q \times 2$ tensor and the optimal approximation of a pair of bilinear forms. In: J. W. de Bakker, J. van Leeuwen (Eds.), *Automata, languages and programming*. Lecture Notes in Computer Science (Vol. 85, pp. 98–108). New York: Springer.

Bro, R., & Smilde, A. K. (2003). Centering and scaling in component analysis. *Journal of Chemometrics, 17*, 16–33.

Carroll, J. D., & Chang, J. J. (1970) Analysis of individual differences in multidimensional scaling via an *n*-way generalization of Eckart-Young decomposition. *Psychometrika, 35*, 283–319.

De Silva, V., & Lim, L.-H. (2008). Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications, 30*, 1084–1127.

Domanov, I., & De Lathauver, L. (2013a). On the uniqueness of the Canonical Polyadic Decomposition of third-order tensors – part I: basic results and uniqueness of one factor matrix. *SIAM Journal on Matrix Analysis and Applications, 34*, 855–875.

Domanov, I., & De Lathauver, L. (2013b). On the uniqueness of the Canonical Polyadic Decomposition of third-order tensors – part II: uniqueness of the overall decomposition. *SIAM Journal on Matrix Analysis and Applications, 34*, 876–903.

Giordani, P., Kiers, H. A. L., & Del Ferraro, M. A. (2014). Three-way component analysis using the R package ThreeWay. *Journal of Statistical Software, 57*(7), 1–23. http://www.jstatsoft.org/article/view/v057i07. Cited January 29, 2016.

Giordani, P., & Rocci, R. (2013a). Candecomp/Parafac via the Lasso. *Psychometrika, 78*, 669–684.

Giordani, P., & Rocci, R. (2013b). Candecomp/Parafac with ridge regularization. *Chemometrics and Intelligent Laboratory Systems, 129*, 3–9.

Giordani, P., & Rocci, R. (2016). Candecomp/Parafac with SVD penalization. Submitted.

Harshman, R. A. (1970). Foundations of the Parafac procedure: Models and conditions for an 'explanatory' multimodal factor analysis. *UCLA Working Papers in Phonetics, 16*, 1–84.

Harshman, R. A., & Lundy, M. E. (1984). Data preprocessing and the extended PARAFAC model. In H. G. Law, C. W. Snyder, J. A. Hattie, & R. P. McDonald (Eds.), *Research methods for multimode data analysis* (pp. 216–284). New York: Praeger.

Jiang, T., & Sidiropoulos N. D. (2004). Kruskal's permutation lemma and the identification of Candecomp/Parafac and bilinear models with constant modulus constraints. *IEEE Transactions on Signal Processing, 52*, 2625–2636.

Kiers, H. A. L., & Van Mechelen, I. (2001). Three-way component analysis: Principles and illustrative application. *Psychological Methods, 6*, 84–110.

Krijnen, W. P., Dijkstra, T. K., & Stegeman, A. (2008). On the non-existence of optimal solutions and the occurrence of "degeneracy" in the Candecomp/Parafac model. *Psychometrika, 73*, 431–439.

Kroonenberg, P. M. (1996). 3WAYPACK User's Manual. http://three-mode.leidenuniv.nl/document/programs.htm#3WayPack. Cited January 29, 2016.

Kroonenberg, P. M. (2008). *Applied multiway data analysis*. Hoboken: Wiley.

Kruskal, J. B. (1977). Three-way arrays: Rank and uniqueness of trilinear decompositions, with applications to arithmetic complexity and statistics. *Linear Algebra and Its Applications, 18*, 95–138.

Kruskal, J. B., Harshman, R. A., & Lundy, M. E. (1989). How 3-MFA data can cause degenerate PARAFAC solutions, among other relationships. In: R. Coppi & S. Bolasco (Eds.), *Multiway data analysis* (pp. 115–122). Amsterdam: Elsevier.

Lim, L.-K., Comon, P. (2009). Nonnegative approximations of nonnegative tensors. *Journal of Chemometrics, 23*, 432–441.

Lundy, M. E., Harshman, R. A., & Kruskal, J. B. (1989). A two stage procedure incorporating good features of both trilinear and quadrilinear models. In: R. Coppi, & S. Bolasco (Eds.), *Multiway data analysis* (pp. 123–130). Amsterdam: Elsevier.

Mitchell, B. C., & Burdick, D. S. (1994). Slowly converging Parafac sequences: Swamps and two-factor degeneracies. *Journal of Chemometrics, 8*, 155–168.

Paatero, P. (2000). Construction and analysis of degenerate Parafac models. *Journal of Chemometrics, 14*, 285–299.

Rocci, R., & Giordani, P. (2010). A weak degeneracy revealing decomposition for the CANDECOMP/PARAFAC model. *Journal of Chemometrics, 24*, 57–66.

Stegeman, A. (2006). Degeneracy in Candecomp/Parafac explained for $p \times p \times 2$ arrays of rank $p + 1$ or higher. *Psychometrika, 71*, 483–501.

Stegeman, A. (2007). Degeneracy in Candecomp/Parafac and Indscal explained for several three-sliced arrays with a two-valued typical rank. *Psychometrika, 72*, 601–619.

Stegeman, A. (2008). Low-rank approximation of generic $p \times q \times 2$ arrays and diverging components in the Candecomp/Parafac model. *SIAM Journal on Matrix Analysis and Applications, 30*, 988–1007.

Stegeman, A. (2009a). On uniqueness conditions for Candecomp/Parafac and Indscal with full column rank in one mode. *Linear Algebra and Its Applications, 431*, 211–227.

Stegeman, A. (2009b). Using the Simultaneous Generalized Schur Decomposition as a Candecomp/Parafac algorithm for ill-conditioned data. *Journal of Chemometrics, 23*, 385–392.

Stegeman, A. (2012). Candecomp/Parafac: From diverging components to a decomposition in block terms. *SIAM Journal on Matrix Analysis and Applications, 30*, 1614–1638.

Stegeman, A. (2013). A three-way Jordan canonical form as limit of low-rank tensor approximations. *SIAM Journal on Matrix Analysis and Applications, 34*, 624–650.

Stegeman, A. (2014). Finding the limit of diverging components in three-way Candecomp/Parafac - a demonstration of its practical merits. *Computational Statistics and Data Analysis, 75*, 203–216.

Stegeman, A., & De Lathauwer, L. (2009). A method to avoid diverging components in the Candecomp/Parafac model for generic $I \times J \times 2$ arrays. *SIAM Journal on Matrix Analysis and Applications, 30*, 1614–1638.

Stegeman, A., ten Berge, J. M. F., & De Lathauwer, L. (2006). Sufficient conditions for uniqueness in Candecomp/Parafac and Indscal with random component matrices. *Psychometrika, 71* 219–229.

ten Berge, J. M. F., Kiers, H. A. L., & De Leeuw, J. (1988). Explicit Candecomp/Parafac solutions for a contrived $2 \times 2 \times 2$ array of rank three. *Psychometrika, 53*, 579–584.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, 58*, 267–288.

Tomasi, G., & Bro, R. (2006). A comparison of algorithms for fitting the PARAFAC model. *Computational Statistics and Data Analysis, 50*, 1700–1734.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika, 31*, 279–311.

Growth Curve Modeling for Nonnormal Data: A Two-Stage Robust Approach Versus a Semiparametric Bayesian Approach

Xin Tong and Zijun Ke

Abstract Growth curve models are often used to investigate growth and change phenomena in social, behavioral, and educational sciences and are one of the fundamental tools for dealing with longitudinal data. Many studies have demonstrated that normally distributed data in practice are rather an exception, especially when data are collected longitudinally. Estimating a model without considering the nonnormality of data may lead to inefficient or even incorrect parameter estimates, or misleading statistical inferences. Therefore, robust methods become important in growth curve modeling. Among the existing robust methods, the two-stage robust approach from the frequentist perspective and the semiparametric Bayesian approach from the Bayesian perspective are promising. We propose to use these two approaches for growth curve modeling when the nonnormality is suspected. An example about the development of mathematical abilities is used to illustrate the application of the two approaches, using school children's Peabody Individual Achievement Test mathematical test scores from the National Longitudinal Survey of Youth 1997 Cohort.

Keywords Growth curve modeling • Robust methods • Semiparametric Bayesian methods • Nonnormality

1 Introduction

Growth curve modeling is one of the most frequently used analytic techniques for longitudinal data analysis with repeated measures because it can directly analyze the intraindividual change over time and interindividual differences in intraindividual change (e.g., McArdle 1988; Meredith & Tisak, 1990). Growth curve analysis is

X. Tong (✉)
University of Virginia, Charlottesville, VA 22904, USA
e-mail: xtong@virginia.edu

Z. Ke
Sun Yat-Sen University, Guangzhou, Guangdong 510275, China
e-mail: keziyun@mail.sysu.edu.cn

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer
Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_17

widely used in social, behavioral, and educational sciences to obtain a description of the mean growth in a population over a specific period of time. Individual variations around the mean growth curve are due to random effects and intraindividual measurement errors. Traditional growth curve analysis typically assumes that the random effects and intraindividual measurement errors are normally distributed. Although the normality assumption makes growth curve models easy to estimate, empirical data usually violate such an assumption. After investigating 440 large scale data sets, Micceri (1989) concluded with an analogy between the existence of normal data and the existence of a unicorn. Practically, data often have longer-than-normal tails and/or outliers. Ignoring the nonnormality of data may lead to unreliable parameter estimates, their associated standard errors estimates, and misleading statistical inferences (see, e.g., Maronna, Martin & Yohai, 2006).

Researchers have become more and more keenly aware of the large influence that nonnormality has upon model estimation (e.g., Hampel, Ronchetti, Rousseeuw & Stahel, 1986; Huber 1981). Some routine methods have been adopted, such as transforming the data so that they are close to being normally distributed, or deleting the outliers prior to fitting a model. However, data transformation can make the interpretation of the model estimation results complicated. Simply deleting outliers may lead the resulting inferences fail to reflect uncertainty in the exclusion process and reduce efficiency (e.g., Lange, Little & Taylor, 1989). Moreover, diagnostics of multivariate outliers in a growth curve model are challenging tasks. High dimensional outliers can be well hidden when the univariate outlier detection methods are used, and are difficult or impossible to identify from coordinate plots of observed data (Hardin & Rocke, 2005). Although various multivariate outlier diagnostic methods have been developed (e.g., Filzmoser 2005; Peña & Prieto, 2001; Yuan & Zhang, 2012a), their detection accuracies are not ideal. Alternatively, researchers have developed what are called robust methods aiming to provide reliable parameter estimates and inferences when the normality assumption is violated.

The ideas of current robust methods falls into two categories. One is to assign a weight to each case according to its distance from the center of the majority of the data, so that extreme cases are downweighted (e.g., Yuan, Bentler & Chan, 2004; Zhong & Yuan, 2010). A few studies have directly discussed this type of robust methods in growth curve analysis. For example, Pendergast and Broffitt (1985) and Singer and Sen (1986) proposed robust estimators based on M-methods for growth curve models with elliptically symmetric errors, and Silvapulle (1992) further extended the M-method to allow asymmetric errors for growth curve analysis. Yuan and Zhang (2012b) developed a two-stage robust procedure for structural equation modeling with nonnormal missing data and applied the procedure to growth curve modeling. Among these methods, the two-stage robust approach is most appealing because it is more stable in small samples and is preferred when the model is not built on solid substantive theory (Zhong & Yuan, 2011). The other category is to assume that the random effects and measurement errors follow certain nonnormal distributions, e.g., t distribution or a mixture of normal distributions. Tong and Zhang (2012) and Zhang, Lai, Lu & Tong (2013) suggested modeling heavy-tailed

data and outliers in growth curve modeling using Student's t distributions and provided online software to conduct the robust analysis. Growth mixture models, first introduced by Muthén and Shedden (1999), provide another useful approach to remedy the nonnormality problem. They assume that individuals can be grouped into a finite number of classes having distinct growth trajectories. Although growth mixture models are very flexible, some difficult issues, including choice of the number of latent classes and selection of growth curve models within each class, have to be tackled. Such issues are automatically resolved by semiparametric Bayesian methods, sometimes referred to as nonparametric Bayesian methods (e.g., Müller & Quintana, 2004), in which the growth trajectories and intraindividual measurement errors are viewed as from random unknown distributions generated from the Dirichlet process. Semiparametric Bayesian method has also been proved to outperform the robust method by using Student's t distributions since Student's t distribution has a parametric form and thus has a restriction on the data distribution (Tong 2014).

Because the two-stage robust approach and the semiparametric Bayesian approach are the most promising method in each category, respectively, and they are also the most promising method from the frequentist and Bayesian perspectives, separately, we propose to use the two approaches to relax the normality assumption in traditional growth curve analysis. In this article, we review the traditional growth curve models and introduce the two robust approaches. The performance of the traditional method and the two robust approaches are then compared by analyzing a simulated dataset with multivariate outliers. The application of the two robust approaches is illustrated through an example with the Peabody Individual Achievement Test math data from the National Longitudinal Survey of Youth 1997 Cohort (Bureau of Labor Statistics, U.S. Department of Labor 2005). We end the article with concluding comments and recommendations.

2 Two Robust Approaches

2.1 Growth Curve Models

Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$ be a $T \times 1$ random vector and y_{ij} be an observation for individual i at time j ($i = 1, \ldots, N; j = 1, \ldots, T$). N is the sample size and T is the total number of measurement occasions. A typical form of growth curve models can be expressed as

$$\mathbf{y}_i = \boldsymbol{\Lambda}\mathbf{b}_i + \mathbf{e}_i,$$
$$\mathbf{b}_i = \boldsymbol{\beta} + \mathbf{u}_i,$$

where $\boldsymbol{\Lambda}$ is a $T \times q$ factor loading matrix determining the growth trajectories, $\mathbf{b}_i$ is a $q \times 1$ vector of random effects, and $\mathbf{e}_i$ is a vector of intraindividual measurement

errors. The vector of random effects $\mathbf{b}_i$ varies for each individual, and its mean, $\boldsymbol{\beta}$, represents the fixed effects. The residual vector $\mathbf{u}_i$ represents the random component of $\mathbf{b}_i$.

Traditional growth curve models typically assume that both $\mathbf{e}_i$ and $\mathbf{u}_i$ follow multivariate normal distributions such that $\mathbf{e}_i \sim MN_T(\mathbf{0}, \boldsymbol{\Phi})$ and $\mathbf{u}_i \sim MN_q(\mathbf{0}, \boldsymbol{\Psi})$, where MN denotes a multivariate normal distribution and the subscript denotes its dimension. The $T \times T$ matrix $\boldsymbol{\Phi}$ and the $q \times q$ matrix $\boldsymbol{\Psi}$ represent the covariance matrices of $\mathbf{e}_i$ and $\mathbf{u}_i$, respectively. For general growth curve models, the intraindividual measurement error structure is usually simplified to $\boldsymbol{\Phi} = \sigma_e^2 \mathbf{I}$ where σ_e^2 is a scalar parameter. By this simplification, we assume the uncorrelatedness of measurement errors and the homogeneity of error variances across time. Given the current specification of $\mathbf{u}_i$, $\mathbf{b}_i \sim MN_q(\boldsymbol{\beta}, \boldsymbol{\Psi})$.

Special forms of growth curve models can be derived from the preceding form. For example, if

$$\boldsymbol{\Lambda} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & T-1 \end{pmatrix}, \mathbf{b}_i = \begin{pmatrix} L_i \\ S_i \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_L \\ \beta_S \end{pmatrix}, \text{and } \boldsymbol{\Psi} = \begin{pmatrix} \sigma_L^2 & \sigma_{LS} \\ \sigma_{LS} & \sigma_S^2 \end{pmatrix},$$

the model represents a linear growth curve model with random intercept (initial level) L_i and random slope (rate of change) S_i. The average intercept and slope across all individuals are β_L and β_S, respectively. In $\boldsymbol{\Psi}$, σ_L^2 and σ_S^2 represent the variability (or interindividual differences) around the mean intercept and the mean slope, respectively, and σ_{LS} represents the covariance between the latent intercept and slope.

In sum, growth curve modeling is a longitudinal analytic technique to estimate growth trajectories over a period of time. The relative standing of an individual at each time is modeled as a function of an underlying growth process, with the best parameter values for that growth process being fitted to the individual. Thus, growth curve modeling can be used to investigate systematic change over time ($\boldsymbol{\beta}$) and interindividual variability in this change ($\boldsymbol{\Psi}$).

2.2 Two-stage Robust Approach

In this section, we review the two-stage robust method developed by Yuan and Bentler (1998).

In the first stage of this method, the saturated mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ of $\mathbf{y}_i$ are estimated by the weighted averages

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{N} w_{i1} \mathbf{y}_i}{\sum_{i=1}^{N} w_{i1}}$$

and

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N} w_{i2}(\mathbf{y}_i - \hat{\boldsymbol{\mu}})(\mathbf{y}_i - \hat{\boldsymbol{\mu}})',$$

respectively, where $w_{i1} = w_1(d_i)$ and $w_{i2} = w_2(d_i)$ are the individual-level weights, and d_i is the Mahalanobis distance, defined by $d_i^2 = d^2(\mathbf{y}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{y}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu})$. A large d_i corresponds to small w_{i1} and w_{i2}. Different weight functions may lead to different estimates. In particular, $\omega_1(d_i) = \omega_2(d_i) = 1$ correspond to normal-based maximum likelihood estimations. In our study, we choose Huber-type weights because they tend to yield more efficient parameter estimates than other weight functions and can effectively control the influence of heavy tails and outliers in real data (e.g., Yuan, Bentler & Chan, 2004).

In the second stage, the robust estimates of the mean vector and covariance matrix are fitted by a structural equation model to find model parameter estimates, standard errors, and related test statistics. Let $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ be the structural model satisfying $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents all the free parameters in the model. The estimates $\hat{\boldsymbol{\theta}}$ are obtained by minimizing

$$F_{ML}(\boldsymbol{\theta}) = [\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\theta})]' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) [\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\theta})] + tr\left[\hat{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\right] - log\left|\hat{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\right| - T.$$

The covariance matrix of $\hat{\boldsymbol{\theta}}$ is consistently estimated by a sandwich-type standard error estimator.

This two-stage robust approach has been further extended to handle missing data as well (Yuan & Zhang, 2012b).

2.3 Semiparametric Bayesian Approach

Since constraining inference to a specific parametric form may limit the scope and type of inferences in many situations, there is growing interest in the use of semiparametric Bayesian methods based on distributions over spaces of distributions. A typical motivation is that one is unwilling to make somewhat arbitrary and unverified assumptions for the latent variables or the error distributions as in the parametric modeling. Unlike typical classical nonparametric methods such as rank and permutation tests, semiparametric Bayesian methods can provide full probability models for the data-generating process and provide posterior distributions of model parameters.

Because the nonnormality of a growth curve model may come from two resources: the measurement errors $\mathbf{e}_i$ and the random components $\mathbf{u}_i$ (see Pinheiro, Liu & Wu, 2001), we assume either one or both of them follow certain nonnormal distributions. Within the semiparametric Bayesian scope, the traditional parametric distributions of $\mathbf{e}_i$ and $\mathbf{u}_i$ are replaced by

$$\mathbf{e}_i \sim G_e,$$

and

$$\mathbf{u}_i \sim G_u,$$

separately, where G_e and G_u are unknown distribution functions. Dirichlet process (DP) (Ferguson 1973, 1974), a distribution over distributions, is used as a prior for the unknown distributions G_e and G_u. The Dirichlet process generates random distributions and is characterized by two hyperparameters, α and G_0. G_0 is a base distribution, which represents the central or "mean" distribution in the distribution space, while the precision parameter α governs how close realizations of G are to G_0. Ferguson (1973) pointed out that the Dirichlet process is a conjugate prior, and has two desirable properties: (1) its support is sufficiently large, and (2) the posterior distribution is analytically manageable.

As mentioned above, either one or both of $\mathbf{e}_i$ and $\mathbf{u}_i$ can be modeled semiparametrically, and thus three types of robust distributional models are proposed. In this article, we focus on a type of model with $\mathbf{e}_i \sim G_e$, $G_e \sim DP$ but $\mathbf{u}_i$ is kept to follow $MN_q(\mathbf{0}, \boldsymbol{\Psi})$. It is named as Semi-N distributional model in Tong (2014). The other two types of distributional models can be estimated similarly. In the Semi-N distributional model, $\mathbf{e}_i$ follows an unknown distribution G_e with a Dirichlet process prior. Because the distribution of $\mathbf{e}_i$ is continuous, G_e is written as a mixture with respect to a mixing measure with the Dirichlet process prior. Therefore, the distribution of $\mathbf{e}_i$ can be a mixture of multivariate normal distributions. We can obtain its distribution by the truncated stick-breaking construction (e.g., Lunn, Jackson, Best, Thomas & Spiegelhalter, 2013; Sethuraman 1994). Assume that the data can be represented by a maximum of C possible mixture components. For $q_1, q_2, \ldots, q_C \sim Beta(1, \alpha)$, define

$$\begin{aligned}
p_1^{'} &= q_1, \\
p_2^{'} &= (1-q_1)q_2, \\
p_3^{'} &= (1-q_1)(1-q_2)q_3, \\
&\vdots \\
p_C^{'} &= (1-q_1)\cdots(1-q_{C-1})q_C,
\end{aligned}$$

with the recursive $p_k^{'} = q_k \prod_{j=1}^{k-1}(1-q_j)$. Then, we can obtain the mixing proportion p_k by

$$p_k = \frac{p_k^{'}}{\sum_{j=1}^{C} p_j^{'}},$$

to satisfy that $\sum_{k=1}^{C} p_k = 1$. Thus, the unknown distribution G_e can be constructed below, which is a mixture of multivariate normal distributions.

$$G_e = \begin{cases} MN(\boldsymbol{\mu}_e^{(1)}, \boldsymbol{\Phi}^{(1)}), & p = p_1 \\ MN(\boldsymbol{\mu}_e^{(2)}, \boldsymbol{\Phi}^{(2)}), & p = p_2 \\ \vdots & \vdots \\ MN(\boldsymbol{\mu}_e^{(C)}, \boldsymbol{\Phi}^{(C)}), & p = p_C \end{cases},$$

where $\boldsymbol{\mu}_e^{(k)}$ and $\boldsymbol{\Phi}^{(k)}, k = 1, \ldots, C$ are parameters of the multivariate normal distribution in the kth component. Because the mean of intraindividual measurement errors $\mathbf{e}_i$ should be $\mathbf{0}$, we let $\boldsymbol{\mu}_e^{(k)} = \mathbf{0}$. For the covariance matrices of the mixture components, $\boldsymbol{\Phi}^{(k)}$, inverse Wishart priors $p(\boldsymbol{\Phi}^{(k)}) = IW(n_0, W_0)$ are used, where n_0 and W_0 are hyperparameters.

In general, G_e can be a multimodal distribution centered at $\mathbf{0}$. The measurement error for the ith individual, $\mathbf{e}_i$, comes from $MN(\mathbf{0}, \boldsymbol{\Phi}^{(k)})$ with the probability p_k. We should point out that the measurement errors for different individuals may be drawn from the same component. For example, it is likely that both $\mathbf{e}_1$ and $\mathbf{e}_2$ follow $MN(\mathbf{0}, \boldsymbol{\Phi}^{(1)})$. It is also quite possible that none of $\mathbf{e}_1, \ldots, \mathbf{e}_N$ are from some components. If $\mathbf{e}_i, i = 1, \ldots, N$ are from K_e different distributions among $MN(\mathbf{0}, \boldsymbol{\Phi}^{(k)}), k = 1, \ldots, C$, K_e is called the number of clusters for $\mathbf{e}_i$. Clearly, $K_e \leq C$, and within each cluster, $\mathbf{e}_i$s come from the same distribution.

Recall that in the traditional growth curve model, $\boldsymbol{\beta}, \boldsymbol{\Phi}$, and $\boldsymbol{\Psi}$ are the model parameters. Here in the Semi-N model, $\boldsymbol{\beta}$ and $\boldsymbol{\Psi}$ are still model parameters and can be estimated in the same way. However, instead of estimating $\boldsymbol{\Phi}$ as in the traditional model, we obtain $\mathbf{e}_i$ and K_e. The estimate of K_e indicates the heterogeneity of intraindividual measurement errors $\mathbf{e}_i$. With a larger value of K_e, we are more confident to conclude that different individuals' measurement errors are distributed differently. To obtain an estimate of $\boldsymbol{\Phi}$ (the covariance matrix of $\mathbf{e}_i$), we let $\mathbf{e}_{i(s)}, i = 1, \ldots, N$ be the observations of $\mathbf{e}_i$ simulated from the posterior distribution in the sth Gibbs sampler iteration, and let $\boldsymbol{\Phi}_{(s)}$ be the corresponding sample covariance matrix. An estimate of $\boldsymbol{\Phi}$ can be taken as the mean of $\boldsymbol{\Phi}_{(s)}$, averaging over all the Gibbs sampler iterations after the burn-in period.

2.4 An Artificial Dataset with Multivariate Outliers to Illustrate the Necessity of the two Robust Approaches Over the Traditional Method

As discussed previously, although the traditional normal-based maximum likelihood (NML) method is widely used for growth curve modeling, it can be deficient because it assumes the normality of data while practical data often violate this assumption. We provide a simulated example in this section to compare the performance of the traditional method to those of the two robust approaches. Two datasets are generated. Dataset 1 (D1), including observations for 100 individuals at 4 time

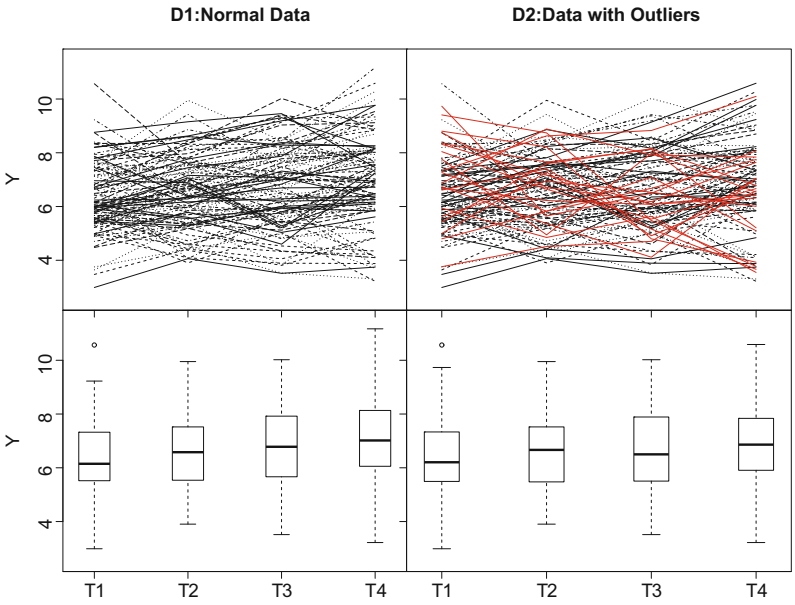


Fig. 1 The trajectory plots and boxplots of two simulated datasets

Table 1 Statistical significance of the latent slope β_S estimates from NML and the two robust approaches

	D1: Normal data	D2: Data with outliers
NML	Significant	Non-significant
Two-stage	Significant	Significant
Semiparametric Bayesian	Significant	Significant

points, is generated from a traditional linear growth curve model with normal assumptions. The latent slope β_S of the overall trajectory is positive. Dataset 2 (D2) is generated by randomly replacing observations for 20 individuals in D1 with multivariate outliers. In particular, the observations for these 20 individuals are generated from a distinct linear growth curve model with slightly larger latent intercept, negative latent slope, and larger intraindividual measurement errors. The trajectory plots and boxplots of D1 and D2 are displayed in Fig. 1. The trajectories for the 20 multivariate outliers in D2 are marked in red. Eyeball examination on those plots fails to locate any suspected outliers, indicating that univariate outlier diagnostic methods risk failure of detecting multivariate outliers, and in this situation, we are vulnerable to outliers that may substantially distort our inferential results.

We fit a linear growth curve model to the two datasets, using NML as well as the two robust approaches, and compare the latent slope β_S estimates. The statistical significance of β_S estimates under different analyses are given in Table 1. For D1,

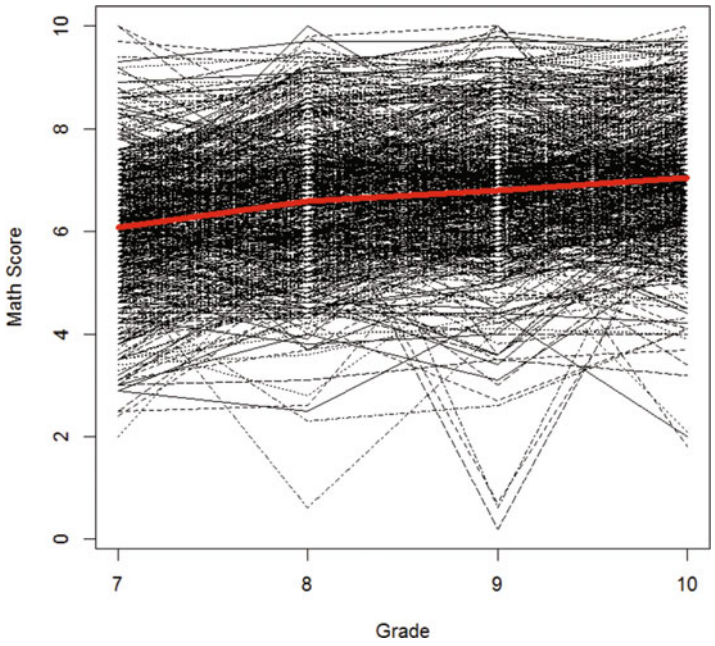


Fig. 2 A collection of individual trajectories for the PIAT math data from NLSY97. 512 school children are measured at 4 occasions

all three methods provide the same results: the latent slope is significantly different from 0. However, for D2, only the two robust approaches are not influenced by multivariate outliers and can still have a significant test result of β_S. This simulated example shows that the two robust approaches perform as well as NML when data are normal, and can provide more reliable inferential results than NML when data contain outliers. This is consistent with previous studies (e.g., Zhong & Yuan, 2011; Tong 2014).

3 An Empirical Example

To demonstrate the application of the two robust approaches, we investigate a subset of data from the National Longitudinal Survey of Youth 1997 (NLSY97) Cohort (Bureau of Labor Statistics, U.S. Department of Labor 2005). In the study, school children's Peabody Individual Achievement Test (PIAT) mathematics scores were collected yearly from the 7th grade to the 10th grade. The individuals' trajectory plot (Fig. 2) suggests a linear growth pattern for the development of math abilities. From the descriptive statistics of the data (Table 2), we notice that the skewness and kurtosis of the data at grades 9 and 10 are significantly different from those of normal distributions. Moreover, the boxplot (Fig. 3) also indicates that there are

Table 2 Descriptive statistics of the PIAT math data from NLSY97

Grade	Mean	s.d.	Skewness	Kurtosis
7	6.071	1.312	−0.110	3.392
8	6.590	1.392	−0.168	3.336
9	6.796	1.419	−0.564*	4.814*
10	7.044	1.325	−0.344*	3.708*

Note The "*" sign indicates that the corresponding statistic is significantly different from that of a normal distribution. The significance of skewness is tested through the D'Agostino test, and the significance of kurtosis is tested through the Anscombe-Glynn test
s.d. = standard deviation

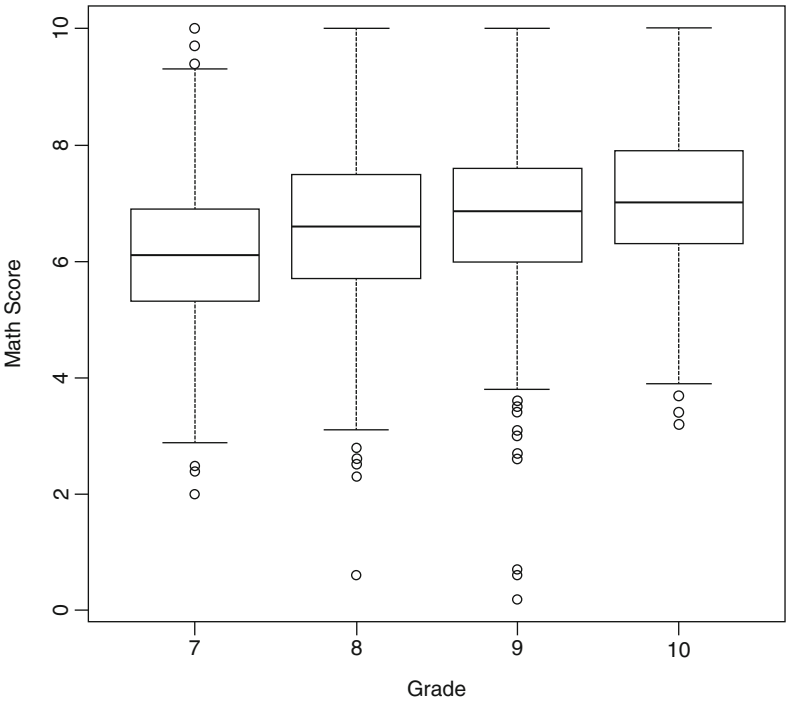


Fig. 3 Boxplot for the PIAT math data from NLSY97. Circles represent potential outliers

potential outliers and the PIAT math scores at each year are skewed to the left. Thus, it is reasonable to consider the data as nonnormal. As a consequence, we will use this dataset to illustrate the application of the robust methods.

A linear growth curve model is fitted to the data and three methods are used for model estimation, including NML, the two-stage robust method, and the semiparametric Bayesian method. The parameter estimates are given in Table 3. The estimates of the latent intercept and slope from the three methods are similar.

Table 3 Parameter estimates from the traditional method as well as the two robust approaches

	NML			Two-stage			Semi-N		
	Est.	CI.L	CI.U	Est.	CI.L	CI.U	Est.	CI.L	CI.U
β_L	6.157	6.047	6.269	6.182	6.073	6.291	6.168	6.058	6.276
β_S	0.312	0.275	0.349	0.319	0.288	0.350	0.314	0.280	0.349
σ_L^2	1.125	0.937	1.337	0.990	0.775	1.205	1.153	0.974	1.357
σ_S^2	0.035	0.024	0.049	0.001	−0.024	0.025	0.040	0.028	0.055
σ_{LS}	−0.034	−0.081	0.009	0.002	−0.050	0.053	−0.049	−0.093	−0.009
σ_e^2	0.748	0.694	0.806	0.580	0.444	0.715	0.737	0.698	0.781

Note Est = estimate; CI.L = lower limit of the 95 % confidence/credible intervals; CI.U = upper limit of the 95 % confidence/credible intervals

The average initial mathematical ability at grade 7 is about 6.2 with an average growth rate of 0.3 from grade 7 to grade 10. By downweighting the potential outliers, variance and covariance estimates from the two-stage method are smaller than those from the other methods. In the two-stage robust method, σ_L^2 estimate is significant, indicating that there are interindividual differences in the initial ability. However, both σ_S^2 and σ_{LS} estimates are not statistically significant, indicating that based on the two-stage method, we have no evidence to deny the claim that the math ability growth rates are the same for all children, regardless of their initial math abilities. For the NML and the semiparametric Bayesian methods, the 95 % confidence/credible intervals of σ_L^2 and σ_S^2 suggest that there are significant interindividual differences in both initial ability and the rate of change. Contrary to the traditional growth curve model with normal assumptions which fails to detect significant estimated σ_{LS}, the robust Semi-N distributional model reports significant negative association between initial math abilities and math ability growth rates. Specifically, the robust Semi-N distributional model suggests that children initially with lower math abilities exhibited higher growth rates in their math abilities from grade 7 to 10. The contradictory results between the traditional model and the Semi-N model are likely related to the width of the estimated confidence/credible intervals. In the presence of nonnormality, the Semi-N distributional model is more efficient, and thus the credible intervals for the Semi-N model are likely to be narrower than the corresponding confidence intervals for the traditional model, which is exactly the case in this example. Given that the correlation between the latent intercept and slope parameters across sessions are interested in many studies (e.g., Zhang, Davis, Salthouse & Tucker-Drob, 2007), the traditional model is not recommended to use when the nonnormality is suspected.

In this analysis, the prior of the DP precision parameter α is given by $Gamma(2, 2)$, as suggested in Ishwaran (2000). The estimate of α is 1.772, resulting in about 12 different clusters for the distribution of measurement errors.

To sum up, the example here documents evidence favoring the semiparametric Bayesian approach as it detects significant results more often than the traditional method and the two-stage robust approach.

4 Concluding Comments

Growth curve models represent repeated measures of dependent variables as a function of time and other measures. These models have grown in use in social and behavioral research since it was shown that they can be fitted under the structural equation modeling framework (Meredith & Tisak, 1990). The traditional growth curve analysis is based upon the normality assumption of random effects and intraindividual measurement errors. However, practical data in social and behavioral sciences are rarely normal because of an unknown population distribution or data contamination. Without properly handling the nonnormality problem, we may get inefficient or even incorrect parameter estimates in model estimation (e.g., Yuan & Bentler, 2001). Studies to deal with the adverse effects of nonnormality on parameter estimates, standard errors, and test statistics have been carried out in growth curve analysis. This article presented two most promising approaches from the frequentist and the Bayesian perspectives, respectively. The simulated data showed that the two robust approaches can reduce the influence of multivariate outliers and should be adopted when the nonnormality is suspected. A systematic comparison between the two approaches deserves further investigations using simulations.

References

Bureau of Labor Statistics, U.S. Department of Labor. (2005). *National Longitudinal Survey of Youth 1997 cohort, 1997–2003 (rounds 1–7) [computer file]*. OSU, Produced by the National Opinion Research Center, the University of Chicago and distributed by the Center for Human Resource Research, The Ohio State University, Columbus, Ohio.

Ferguson, T. (1973) A bayesian analysis of some nonparametric problems. *The Annals of Statistics, 1*, 209–230.

Ferguson, T. (1974) Prior distributions on spaces of probability measures. *The Annals of Statistics, 2*, 615–629.

Filzmoser, P. (2005) Identification of multivariate outliers: A performance study. *Austrian Journal of Statistics, 34*, 127–138.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986) *Robust statistics: The approach based on influence functions*. New York: Wiley.

Hardin, J., & Rocke, D. M. (2005) The distribution of robust distances. *Journal of Computational and Graphical Statistics, 14*, 928–946.

Huber, P. J. (1981) *Robust statistics*. New York: Wiley.

Ishwaran, H. (2000) Inference for the random effects in bayesian generalized linear mixed models. In American Statistical Association, (ed.), *ASA Proceedings of the Bayesian Statistical Science Section* (pp. 1–10).

Lange, K. L., Little, R. J. A., & Taylor, J. M. G. (1989) Robust statistical modeling uisng the t distribution. *Journal of the Americal Statistical Association, 84(408)*, 881–896.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2013) *The BUGS book: A practical introduction to Bayesian analysis*. Boca Raton, FL: CRC Press.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006) *Robust statistics: Theory and methods*. New York: Wiley.

McArdle, J. J. (1988) Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade, & R. B. Cattell, (eds.), *Handbook of Multivariate Experimental Psychology* (2nd ed., pp. 561–614). New York: Plenum Press.

Meredith, W., & Tisak, J. (1990) Latent curve analysis. *Psychometrika, 55*, 107–122.

Micceri, T. (1989) The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105(1)*, 156–166.

Müller, P., & Quintana, F. A. (2004) Nonparametric bayesian data analysis. *Statistical Science, 19*, 95–110.

Muthén, B., & Shedden, K. (1999) Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics, 55(2)*, 463–469.

Peña, D., & Prieto, F. J. (2001) Multivariate outlier detection and robust covariance matrix estimation (with discussion). *Technometrics, 43*, 286–310.

Pendergast, J. F., & Broffitt, J. D. (1985) Robust estimation in growth curve models. *Communications in Statistics: Theory and Methods, 14*, 1919–1939.

Pinheiro, J. C., Liu, C., & Wu, Y. N. (2001) Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics, 10(2)*, 249–276.

Sethuraman, J. (1994) A constructive definition of dirichlet priors. *Statistica Sinica, 4*, 639–650.

Silvapulle, M. J. (1992) On m-methods in growth curve analysis with asymmetric errors. *Journal of Statistical Planning and Inference, 32(3)*, 303–309.

Singer, J. M., & Sen, P. K. (1986) M-methods in growth curve analysis. *Journal of Statistical Planning and Inference, 13*, 251–261.

Tong, X. (2014) *Robust semiparametric bayesian methods in growth curve modeling*. Unpublished doctoral dissertation, University of Notre Dame, Notre Dame.

Tong, X., & Zhang, Z. (2012) Diagnostics of robust growth curve modeling using student's t distribution. *Multivariate Behavioral Research, 47*, 493–518.

Yuan, K.-H., & Bentler, P. M. (1998) Structural equation modeling with robust covariances. *Sociological Methodology, 28*, 363–396.

Yuan, K.-H., & Bentler, P. M. (2001) Effect of outliers on estimators and tests in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology, 54*, 161–175.

Yuan, K.-H., Bentler, P. M., & Chan, W. (2004) Structural equation modeling with heavy tailed distributions. *Psychometrika, 69*, 421—436.

Yuan, K.-H., & Zhang, Z. (2012a) Structural equation modeling diagnostics using r package semdiag and eqs. *Structural Equation Modeling, 19*, 683–702.

Yuan, K.-H., & Zhang, Z. (2012b) Robust structural equation modeling with missing data and auxiliary variables. *Psychometrika, 77*, 803–826.

Zhang, Z., Davis, H. P., Salthouse, T. A., & Tucker-Drob, E. A. (2007) Correlates of individual, and age-related, differences in short-term learning. *Learning and Individual Differences, 17(3)*, 231–240.

Zhang, Z., Lai, K., Lu, Z., & Tong, X. (2013) Bayesian inference and application of robust growth curve models using student's t distribution. *Structural Equation Modeling, 20*, 47–78.

Zhong, X., & Yuan, K.-H. (2010) Weights. In N. J. Salkind (ed.), *Encyclopedia of research design* (pp. 1617–1620). Thousand Oaks: Sage.

Zhong, X., & Yuan, K.-H. (2011) Bias and efficiency in structural equation modeling: Maximum likelihood versus robust methods. *Multivariate Behavioral Research, 46*, 229–265.

The Specification of Attribute Structures and Its Effects on Classification Accuracy in Diagnostic Test Design

Ren Liu and Anne Corinne Huggins-Manley

Abstract Diagnostic test has gained attention for its potentiality to produce fine-grained information about examinees. The dependency among attributes (i.e. attribute structure) is one of the most important factors affecting diagnostic test design. This article introduces four types of attribute structures and examines the effects of the attribute number, structure and level on classification accuracy and reliability. Results from the study help researchers and practitioners understand factors that affect classification when specifying attributes, and design diagnostic tests that provide accurate information about examinees.

Keywords Diagnostic classification model • Classification accuracy • Attribute structure • Hierarchical diagnostic classification model • Attribute hierarchy method • Test design • Cognitive diagnostic measurement

1 Introduction

Classifying examinees at the skill level is a test outcome desired by many educational practitioners. At the national level, personalized learning is named as a top priority (U.S. Department of Education 2014), which emphasizes tailored instructional improvements that include providing personalized feedback on the strengths and weaknesses of students on specific learning objectives in K-12 assessments. Recent developments in diagnostic test design and diagnostic classification models (DCMs) offer the possibility of gaining information on skill mastery profiles from examinee item responses.

There are three core parts at the design stage of a diagnostic test: the specification of the Q-matrix, the design of the Q-matrix and the specification of the attribute structure. The specification of the Q-matrix refers to the process of designing items that measure attributes, and it requires specific content knowledge. Misspecification of Q-matrix occurs when items that are designed to measure specific attributes do

R. Liu (✉) • A.C. Huggins-Manley
Department of Research and Evaluation Methodology, University of Florida, Gainesville, FL 32611, USA
e-mail: liurenking@ufl.edu

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_18

not achieve that purpose well, and it results in misclassification of examinees as evidenced by Rupp and Templin (2008) and Kunina-Habenicht, Rupp, and Wilhem (2012). The design of the Q-matrix refers to the process of loading items on attributes. Three approaches of Q-matrix design have been proposed including the linear approach, the adjacent approach, and the reachable approach (Liu, Huggins-Manley, Bradshaw 2016). They show that the adjacent approach provides higher classification accuracy in a shorter test and is recommended in future design when attributes form a hierarchy. The specification of attribute structure refers to the process of specifying the relationships amongst attributes, and it precedes the specification and design of the Q-matrix. The specified attribute hierarchies in diagnostic tests are formalizations of these attribute dependencies that are grounded in hypothesized learning trajectories. Although it precedes the Q-matrix specification and design, research has been lacking on how the specification of attribute structure affects classification results.

The purpose of the study is to propose four types of attribute structures and examine their effects on classification accuracy in diagnostic test design. The research questions include how each of the following four facets affects classification accuracy and reliability when holding the others constant: (a) the number of attributes; (b) the type of attribute structure; (c) the number of levels specified in the structure; and (d) the specific level at which an attribute is located in the structure. The answers to these questions will help researchers and practitioners specify attribute structures and design diagnostic tests. The remainder of the manuscript is organized as follows. First we introduce three mainstays of diagnostic test design and focus on the specification of attribute structures through proposing four types of structures. Then, a simulation study is conducted to examine how each of the abovementioned four facets affects classification results. Discussions are given at the end.

2 Specification of Attribute Structures

When we look into students' learning trajectories, we often find that learning is a sequential process, with each step built upon the previous one. This implies that the skills students learn are hierarchical because mastering some skills is a pre-requisite to learning other skills. The resulting hierarchy of attributes, which may be evident in students' responses, should also be specified when we design diagnostic tests.

One study that addressed the specification of attribute structure is Leighton, Gierl, and Hunka (2004), where they proposed four types of attribute structures. However, their specifications are confounded and they did not examine how different attribute structures affect classification results. Thus, before discussing how attribute structure affects classification, we first need to form a clear taxonomy

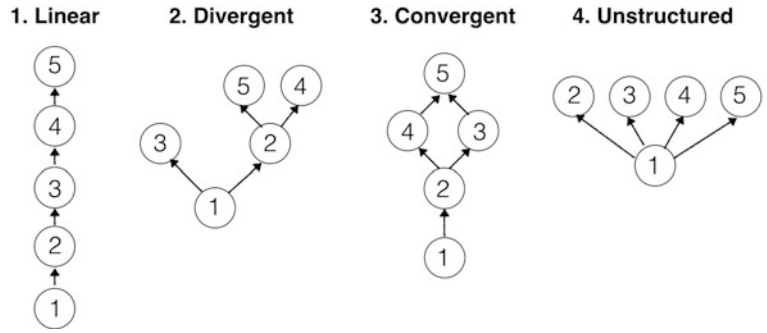


Fig. 1 Linear, divergent, convergent and unstructured hierarchy using five attributes

of the types of structures. In the following paragraphs, we first present the Leighton et al. (2004) taxonomy of attribute structures, and then introduce the four structures that we propose.

Leighton et al. (2004) proposed four types of attribute hierarchies: *linear*, *divergent*, *convergent and unstructured* as illustrated in Fig. 1. In a linear hierarchy, all five attributes are sequentially ordered in one single chain. In a divergent hierarchy, multiple branches diverge from a common parent attribute. In a convergent hierarchy, multiple parent attributes converge to a common attribute. In an unstructured hierarchy, one attribute precedes multiple distinct attributes. However, there are some confounding issues among the four structures in their taxonomy. For example, the unstructured hierarchy can be viewed as a special case of the divergent structure where a parent attribute only produces one level of offspring. Also, the divergent structure can be viewed as a special case of the convergent structure where multiple attributes do not converge to one.

To avoid confounding issues, we propose four types of attribute structures: *linear*, *pyramid*, *inverted pyramid and diamond*, as illustrated in Fig. 2. In a *linear* structure, all attributes are sequentially ordered in one chain, which aligns with Leighton et al. (2004). In our example, examinees who have mastered higher-level $a5$ are expected to have mastered all preceding attributes (i.e., $a1$ to $a4$). In a *pyramid* structure, multiple parent attributes converge to a common child attribute, and one parent can produce at most one offspring. Thus, examinees who have mastered higher-level $a5$ are expected to have mastered one of the four parent attributes (i.e. $a1$ to $a4$). The pyramid structure narrows the scope of the convergent structure, and allocates the mixture of convergent and divergent structure to the diamond structure. In an *inverted pyramid* structure, one parent attribute produces multiple offspring, and a higher-level attribute can only have one parent in the lower level. Thus, examinees who have mastered either one of the higher-level attributes from $a2$ to $a5$ are expected to have mastered the parent attribute (i.e. $a1$). The inverted pyramid structure absorbs the traditional divergent and unstructured hierarchy. The *diamond* structure refers to the combination of two or three structures. In our example, linear structure (formed by $a1$ and $a2$), inverted pyramid (formed by $a2$, $a3$, and

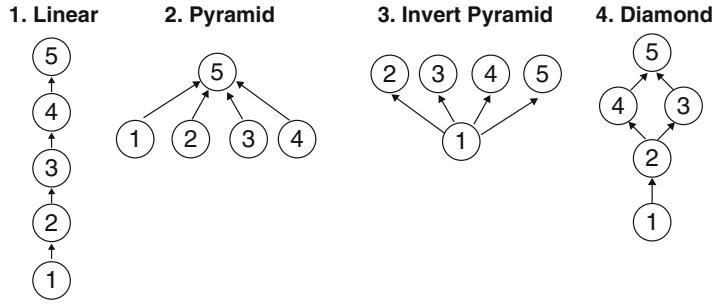


Fig. 2 Linear, pyramid, inverted pyramid and diamond structures using five attributes

$a4$), and pyramid (formed by $a3$, $a4$, and $a5$) co-constructed the diamond structure. Thus, examinees who have mastered $a5$ are expected to have mastered one or more of the preceding attributes (i.e., either attribute 1, 2, and 3; or attribute 1, 2, and 4; or attribute 1, 2, 3, and 4).

The four types of structure we propose offer a solid foundation for us to understand their effects on classification accuracy in the following sections. In this article, to isolate the effects of attribute structure on classification, we held the Q-matrix design constant by applying the same approach to load items on attributes, and assumed all Q-matrices are correctly specified.

3 Method

There are two lines of work on modeling attribute hierarchies: Hierarchical Diagnostic Classification Model (HDCM; Templin & Bradshaw 2014), and Attribute Hierarchy Method (AHM; Leighton et al. 2004). HDCM adapts the saturated Log-linear Cognitive Diagnosis Model (LCDM; Henson, Templin & Willse 2009) in that redundant parameters are reduced to reflect the nested structure of attributes. Suppose an item measuring attributes a and b, where b is nested within a. The item response function under the LCDM for an examinee e on item i is

$$P\left(y_{ei}=1\,|\boldsymbol{a_e}\right)=\frac{\exp\left(\lambda_{i,0}+\lambda_{i,1,\,(a)}a_{ea}+\lambda_{i,1,\,(b)}a_{eb}+\lambda_{i,2,\,(a,b)}a_{ea}a_{eb}\right)}{1+\exp\left(\lambda_{i,0}+\lambda_{i,1,\,(a)}a_{ea}+\lambda_{i,1,\,(b)}a_{eb}\lambda_{i,2,\,(a,b)}a_{ea}a_{eb}\right)}. \quad (1)$$

while the item response function under the HDCM model is

$$P\left(y_{ei}=1\,|\boldsymbol{a_e}\right)=\frac{\exp\left(\lambda_{i,0}+\lambda_{i,1,\,(a)}a_{ea}+\lambda_{i,2,\,(b(a))}a_{ea}a_{eb}\right)}{1+\exp\left(\lambda_{i,0}+\lambda_{i,1,\,(a)}a_{ea}+\lambda_{i,2,\,(b(a))}a_{ea}a_{eb}\right)}. \quad (2)$$

In both equations, $\boldsymbol{\alpha}_c$ denotes the attributes in profile c, $\lambda_{i,0}$ is an intercept parameter, representing the logit of a correct response where the all entries in the examinee's α_c equal 0. $\lambda_{i,1,(a)}$ is the main effect associated with attribute a, $\lambda_{i,1,(b)}$ is the main effect associated with attribute b, and both $\lambda_{i,2,(a,b)}$ in Eq. (1) and $\lambda_{i,2,(b(a))}$ in Eq. (2) represent the two-way interaction effect parameter associated with attribute a and attribute b.

AHM is another line of work in modeling attribute hierarchy, and it is a variation from Tatsuoka's Rule-Space Model (RSM; Tatsuoka 1983 2009). As a pattern classification approach, it is similar to RSM in the way that the observed response patterns (OP) are classified by matching expected response patterns (EP). However, it is different from RSM in the way that attributes are assumed to be dependent to form hierarchies. Therefore, AHM does not identify incorrect "rules", but identifies attribute combinations that are and are not available to examinees. The AHM produces an estimation of the likelihood that OP approximate EP at a given θ. The conditional probability of an examinee e's OP being the same as the EP is modeled as a function of the product of the likelihood of all J slips from 0 to 1 and all K slips from 1 to 0, or

$$P_{e(EP=OP)}\left(\theta_{EP}\right) = \prod_{j=1}^{J} P_{ej}\left(\theta_{EP}\right) \prod_{k=1}^{K} \left(1 - P_{ek}\left(\theta_{EP}\right)\right). \tag{3}$$

The examinee is classified as having the cth set of attributes when $P_{e(EP=OP)}\left(\theta_{EP}\right)$ is the largest. The HDCM has shown to produce higher classification accuracy across different attribute structures as compared to the AHM (Liu & Huggins-Manley 2015); therefore, it is used as the psychometric model in the simulation study.

4 Simulation Study

To explore the effects of attribute structures on classification accuracy, our simulation study manipulated three key factors: the number of attributes, the type of attribute structure, and the number of attribute levels. In total, we specified 12 attribute structures, which are displayed in Fig. 3. The resulted 12 simulation conditions are outlined in Table 1. Specifically, the number of attributes was set at two levels: three and five. The formation of a diamond structure needs at least four attributes; therefore, the 3-attribute specifications were only applied to linear, pyramid and inverted pyramid structures. To examine how the number of attributes affects classification results while holding the type of attribute structure constant, L3 and L5 were specified under the linear structure, P3, P5-1, P5-2 were specified under the pyramid structure, and IP3, IP5-1, IP5-2 were specified under the inverted pyramid structure.

To examine how types of structure affect classification results when holding the attribute number and levels constant, L3 was specified to compare with P3 and IP3. P5-1 and P5-2 were specified to compare with IP5-1 and IP5-2 respectively. L5

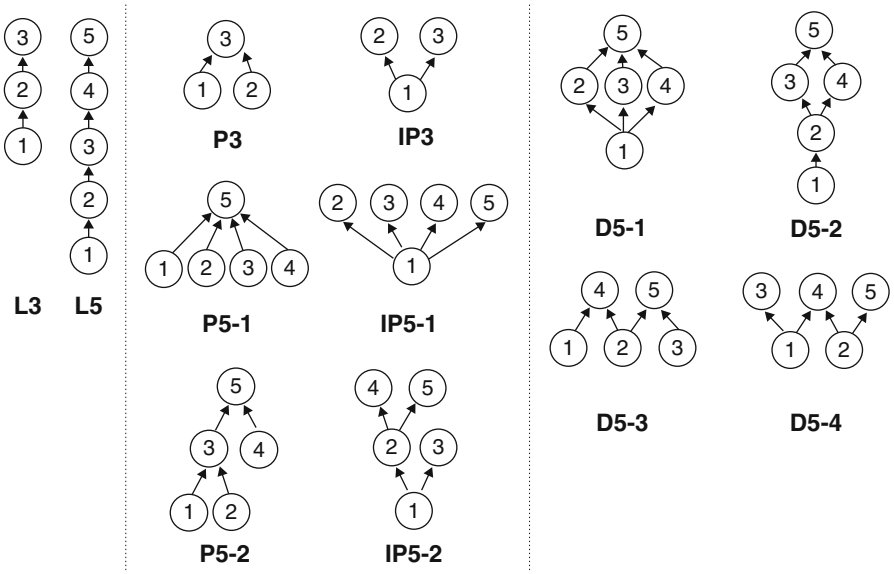


Fig. 3 12 Attribute structures specified in the simulation study

was also compared with conditions in the pyramid, inverted pyramid and diamond structures where five attributes were used. To examine how the number of attribute levels affects classification results when holding the attribute number and type of structure constant, P5-1, IP5-1, D5-1 were specified to compare with P5-2, IP5-2, and D5-2.

The tetrachoric correlations among attributes were fixed at .70, similar to Bradshaw and Templin (2014). 2000 examinees were drawn from a normal distribution, and their attribute profiles were estimated through maximum *a posteriori* (MAP). The item parameters were generated from a uniform distribution, where $P(1\,|\boldsymbol{\alpha}_c = 0)$ and $P(1\,|\boldsymbol{\alpha}_c = 1)$ were generated from $U(0.15, 0.30)$, and $U(0.70, 0.85)$ respectively to represent a mid-quality item bank, similar to de la Torre (2009). We used the adjacent approach to design the Q-matrices (Liu et al. 2016), where eight items were used to measure each attribute five times. Given a small number of items, we expected that the classification accuracy for most profiles would be below .90 across conditions, and it is worth mentioning that the purpose of our study is not to infer from absolute classification results, but rather to infer from relative classification results across different conditions. Simulation studies in previous research (e.g. Madison & Bradshaw 2015) show the classification accuracy hits a ceiling when a large number of items are used. To avoid ceiling effects that can mask factors associated with poor classification accuracy, we purposely used a small number of items. Each condition was replicated 100 times to provide stable results. Three indices of classification accuracy and reliability were used to compare across simulation designs. The accuracy of classification was evaluated by the attribute-

Table 1 Simulation conditions

Structure	Name	Description	Levels
Linear	L3	Three attributes form a linear sequence where $a1$ is at the lowest level and $a3$ is at the highest level	3
	L5	Five attributes form a linear sequence where $a1$ is at the lowest level and $a5$ is at the highest level	5
Pyramid	P3	$a1$ and $a2$ converge to $a3$	2
	P5-1	Four attributes from $a1$ to $a4$ converge to $a5$	2
	P5-2	$a1$ and $a2$ converge to $a3$. $a3$ and $a4$ converge at the second level to produce $a5$	3
Inverted Pyramid	IP3	$a1$ and $a2$ converge to $a3$	2
	IP5-1	$a1$ branches out into four other attributes (i.e. $a2$ to $a5$)	2
	IP5-2	$a1$ branches out into $a2$ and $a3$. $a2$ at the second level branches out into $a4$ and $a5$	3
Diamond	D5-1	$a1$ branches out into $a2$, $a3$, and a4. $a2$, $a3$, and $a4$ at the second level converge to $a5$	3
	D5-2	$a1$ is a linear prerequisite to $a2$. $a2$ at the second level branches out into $a3$ and $a4$. $a3$ and $a4$ at the third level converge into $a5$	4
	D5-3	$a1$ and $a2$ converge to $a4$. $a2$ and $a3$ converge to $a5$	2
	D5-4	$a1$ and $a2$ converge to $a4$, $a1$ also branches out into $a3$, and $a2$ also branches out into $a5$	2

wise classification accuracy (ACA) and the profile classification accuracy (PCA). Each criterion was computed as

$$\text{ACA} = \sum_{e=1}^{N} \sum_{a=1}^{A} \frac{E\left[\widehat{a}_{ea} = a_{ea}\right]}{NA}, \text{ and} \tag{4}$$

$$\text{PCA} = \sum_{e=1}^{N} \frac{E\left[\widehat{a}_{e} = a_{e}\right]}{N}, \tag{5}$$

where N is the number of examinees, $\hat{a}_{ea}$ is the estimated mastery status for examinee e on the ath attribute, and $\hat{a}_e$ is the estimated mastery pattern for examinee e. The reliability of classification is the stability of examinee classification across the iterations, and it was evaluated by the average reliability across all attributes.

Table 2 Attribute-wise classification accuracy

Structure	Name	$\alpha 1$	$\alpha 2$	$\alpha 3$	$\alpha 4$	$\alpha 5$
Linear	L3	0.920	**0.897**	0.960		
	L5	0.948	0.988	0.998	1.000	1.000
Pyramid	P3	**0.872**	**0.824**	0.961		
	P5-1	**0.865**	**0.870**	**0.869**	**0.859**	0.994
	P5-2	0.908	0.907	0.997	0.915	0.998
Inverted pyramid	IP3	**0.890**	**0.881**	**0.876**		
	IP5-1	0.913	0.915	0.913	**0.894**	0.917
	IP5-2	0.901	0.946	0.982	0.983	0.983
Diamond	D5-1	**0.881**	0.961	0.964	0.958	0.995
	D5-2	0.944	0.979	0.997	0.997	0.998
	D5-3	**0.886**	**0.879**	**0.893**	0.995	0.996
	D5-4	**0.885**	0.903	0.979	0.988	0.983

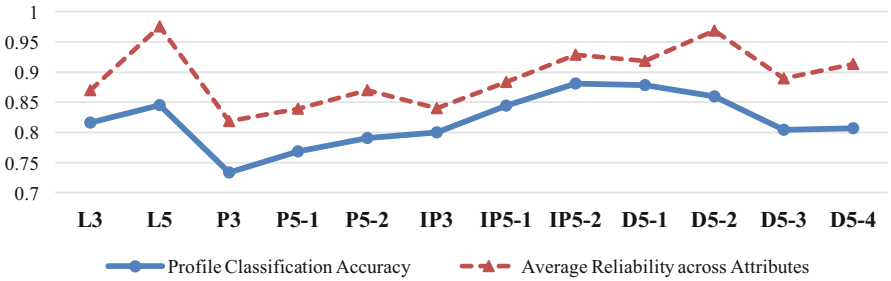


Fig. 4 Profile classification accuracy and average reliability across attributes

5 Results

Results of the simulation study are discussed with a focus on the primary research inquiry: the effects of attribute structure on examinee classification. The ACAs of each attribute are presented in Table 2, the PCA and average reliability across attributes in each condition are illustrated in Fig. 4. We discuss the results in four parts: the effects of attribute numbers on classification, the effects of structure types on classification, the effects of the number of attribute levels on classification, and the effects of level on attribute-wise classification.

5.1 The Effects of Attribute Numbers on Classification

In the linear structure the PCA for L5 was .846, higher than .816 of L3. The ACAs for all attributes in L5 were also higher than those in L3. The ACAs of three attributes in L3 ranged from .897 to .960, while the ACAs of five attributes in L5 ranged from .948 to 1.000. This trend was also observed in the pyramid and inverted

pyramid structures. From P3 to P5-2, the PCA increased from .728 to .772 in the pyramid structure. From IP3 to IP5-2, the PCA also increased from .793 to .870. The ACAs for P5-2 and IP5-2 were all above .90, while the numbers were mostly below .90 for P3 and IP3. The average reliability across attributes in different conditions showed the same trend. L5 produced .976 average reliability while L3 produced .870. P5-2 and IP5-2 produced average reliability of .910 and .929 respectively, while P3 and IP-3 produced .819 and 800, respectively. To summarize, results showed that a larger number of attributes tended to yield high classification accuracy and reliability when holding the type and level of attribute structure constant.

5.2 The Effects of Structure Types on Classification

When the number of attributes was set at three, the PCA for L3, P3 and IP3 was .814, .728, and .794, respectively. The ACAs and average reliability in the three conditions and that the linear structure produced the highest classification accuracy reliability. When the number of attributes was set at five, the PCA for P5-1, P5-2, IP5-1, and IP5-2 was .769, .790, 845, and 881, respectively. It is clear that the inverted pyramid structure produced higher classification accuracy than the pyramid structure when the number of attributes and attribute levels were held constant. We also compare across five-attribute conditions across different structures. The linear structure produced higher classification accuracy than P5-1, but lower than P5-2. The difference between P5-1 and P5-2 is the attribute levels, which will be discussed in the next section. Four conditions under the diamond structure produced PCA higher than .80, and D5-2 produced .890 PCA, the highest among all 12 conditions. To conclude, results showed that the pyramid structure produced lower classification accuracy than the other three structures, but there are other factors affecting the classification in each structure when holding the number of attributes constant.

5.3 The Effects of the Number of Attribute Levels on Classification

When the number of attributes was set at 3, there were only two levels that could be specified. When the number of attributes was fixed at 5, two, three, or four levels could be specified. P5-1 and P5-2 were specified under the pyramid structure with two and three levels respectively, and IP5-1 and IP5-2 were specified under the inverted pyramid structure with two and three levels respectively. P5-1 produced PCA of .768, and the number increased to .790 for P5-2. Similarly, IP5-1 produced PCA of .844 and the number increased to .881 for IP5-2. The ACAs in P5-2 and IP5-2 were all above .90, better than P5-1 and IP5-1. The average reliability for P5-1, P5-2, IP5-1 and IP5-2 were .839, 870, .883 and .929 respectively. Therefore,

under the pyramid and the inverted pyramid structure, introducing more attribute levels within a fixed number of attributes produced higher classification accuracy and reliability. The number of attributes equals the level of attributes in the linear structure; therefore it aligned with our finding that specifying attributes into more levels produced higher classification results under the linear structure. When the attribute structure was diamond, D5-1, D5-2, D5-3 and D5-4 had 3, 4, 2, and 2 levels, respectively. As mentioned in a previous section, D5-2 produced the highest PCA, followed by D5-1, D5-4 and D5-3. The ACAs in D5-2 were from .944 to .998, while there were ACAs below .90 in other three conditions. Results of the average reliability followed the same trend with the PCA. In conclusion, when holding the type of attribute structure and the attribute number constant, specifying attributes into more levels produced higher classification accuracy and reliability.

5.4 The Effects of Level on Attribute-Wise Classification

We also investigated whether higher-level attributes and lower level attributes produced systematically different ACAs. We found that higher-level attributes were often associated with higher classification accuracy than lower-level attributes. For example in P5-1, the four attributes from $a1$ to $a4$ had ACAs less than .90, and $a5$ had .99 ACA. In P5-2, $a1$, $a2$, and $a4$ had ACAs about .91, and $a3$ and $a5$ had .99 ACAs. In D5-1, $a1$ had the only PCA less than .90 among five attributes. In D5-3, the ACAs of the first three attributes were below .90, and numbers were .99 for $a4$ and $a5$. In D5-4, $a1$ and $a2$ had PCA around .89, while $a3$ and $a5$ had PCA around .98.

6 Discussion

This study demonstrated the effects of attribute numbers, structure types and attribute levels on classification accuracy and reliability. The major findings are: (a) when holding the type and number of levels of attribute structure constant, increasing the number of attributes produces higher classification accuracy and reliability; (b) when holding the attribute number constant, the pyramid structure produces lower classification accuracy and reliability than the other three structures; (c) when holding the type of attribute structure and the attribute number constant, specifying attributes into more levels produces higher classification accuracy and reliability; and (d) higher-level attributes produce higher classification results as compared to lower-level attributes.

One of the core rationales behind our findings is that attributes with more arrows directly or indirectly arriving at them have higher classification accuracy. More arrows arriving at one attribute indicates that there is more information from the hierarchical structure about that attribute when the attribute number, structure or

levels is held constant. Specifically, when the attribute hierarchy is specified *a priori*, the number of permissible mastery profiles is limited. For example, in L5 if an examinee has not mastered $a1$, we do not expect that he/she has mastered $a2$. Similarly, if an examinee has not mastered any of the four attributes: $a1$, $a2$, $a3$, or $a4$, we can be certain that he/she has not mastered $a5$. Therefore, more attribute numbers or higher attribute levels may produce higher classification accuracy. It is also the reason why the pyramid structure produces lower accuracy. For example, in P5-1, four attributes $a1$, $a2$, $a3$, and $a4$ have low ACAs (< .87) because they do not have preceding attributes, and this means there is no information from the hierarchical structure contributing to them.

Our taxonomy of attribute structures and results from this study can be used as a guide for researchers and practitioners in specifying attribute structures. We also want to remind test developers that there are other factors that affect the classification results such as Q-matrix specification and design. Although the current study proposed four types of attribute structures and examined how attribute structures affect classification accuracy, it has limitations in several aspects. First, Q-matrix specification is assumed to be correct, and Q-matrix design is fixed in the study. Interested researchers can investigate the interplay among the specification and design of the Q-matrix and the attribute structure. Second, the number of items used in the study is intentionally set to be small. It would be helpful to investigate how the number of items affects classification accuracy and find the point of diminishing returns. Third, the diamond structure is a combination of two or three structures, and studies on the different combinations of attribute structures would be needed in designing diagnostic tests.

References

Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika, 79*(3), 403–425.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115–130.

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.

Kunina-Habenicht, O., Rupp, A. A., & Wilhem, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement, 49*, 59–81.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy model for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*, 205–237.

Liu, R., Huggins-Manley, A. C. (2015). *Validating attribute structures in the attribute hierarchy method for making diagnostic inferences.* Paper presented at the International Meeting of the Psychometric Society in Beijing, China.

Liu, R., Huggins-Manley, A.C., Bradshaw, L. (2016). The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educational and Psychological Measurement.* DOI: 10.1177/0013164416645636

Madison, M. J., & Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the Log-linear cognitive diagnosis model. *Educational and Psychological Measurement, 75*(3), 491–511.

Rupp, A. A., & Templin, J. (2008). Effects of Q-matrix misspecification on parameter estimates and misclassification rates in the DINA model. *Educational and Psychological Measurement, 68*, 78–98.

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

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York: Routledge.

Templin, J., & Bradshaw, L. P. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*, 317–339.

U.S. Department of Education. (2014). *Secretarys final supplemental priorities and definitions for discretionary grant programs*. Retrieved from https://www.federalregister.gov/articles/2014/12/10/2014-28911/secretarys-final-supplemental-priorities-and-definitions-for-discretionary-grant-programs#h-28.

Conditions of Completeness of the Q-Matrix of Tests for Cognitive Diagnosis

Hans-Friedrich Köhn and Chia-Yi Chiu

Abstract The associations between the items of a test based on the cognitive diagnosis framework and the skills required to solve them are documented in the Q-matrix. If the items have skill profiles that allow for the identification of all possible proficiency classes among examinees, then the Q-matrix of the test is said to be complete. An incomplete Q-matrix causes examinees to be assigned to proficiency classes to which they do not belong. Thus, completeness of the Q-matrix is an integral requirement of any cognitively diagnostic test. However, completeness of the Q-matrix is often difficult to establish, especially, for tests with a large number of items involving multiple skills. As an additional complication, completeness is not an intrinsic property of the Q-matrix, but can only be assessed in reference to a specific diagnostic classification model (DCM) supposed to underlie the data—that is, the Q-matrix of a given test can be complete for one model but incomplete for another. For different types of DCMs, conditions of Q-completeness are studied. Rules are derived to determine the completeness of a given Q-matrix.

Keywords Cognitive Diagnosis • Diagnostic Classification Models • Q-Matrix Completeness • Identifiability

1 Introduction

Cognitive diagnosis (CD) in educational assessment (DiBello, Roussos & Stout, 2007; Haberman & von Davier, 2007; Leighton & Gierl, 2007; Rupp, Templin & Henson, 2010) seeks to assign students to proficiency classes that are defined in terms of distinct profiles of (discrete) cognitive skills, called attributes, that a student may have mastered or not.

H.-F. Köhn (✉)
University of Illinois at Urbana-Champaign, Champaign, IL, USA
e-mail: hkoehn@illinois.edu

C.-Y. Chiu
Rutgers, The State University of New Jersey, New Brunswick, NJ, USA
e-mail: chia-yi.chiu@gse.rutgers.edu

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer
Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_19

Educational tests constructed based on the CD framework use items that themselves are characterized by individual profiles that specify which attributes are required to respond correctly to an item. The entire set of these item-attribute associations constitutes the Q-matrix of a test (Tatsuoka 1985). The Q-matrix is an integral component of any test that is based on the CD framework. The Q-matrix must fulfill the requirement that it be complete—that is, it must allow for the identification of all possible proficiency classes among examinees (Chiu, Douglas & Li, 2009). Said differently, an incomplete Q-matrix does not allow for the identification of all proficiency classes thus, causing examinees to be assigned to proficiency classes to which they do not belong. Completeness of the Q-matrix is therefore a key requirement of any CD test.

However, completeness of the Q-matrix is often difficult to establish, especially, for tests with a large number of items involving multiple attributes. As an additional complication, completeness is not an intrinsic property of the Q-matrix, but can only be assessed in reference to a specific diagnostic classification model (DCM) supposed to underlie the data. In other words, the Q-matrix of a given test can be complete for one model, but incomplete for another.

In this article, the results of examining the conditions of Q-completeness for different DCMs are reported. Rules are derived for determining whether a given Q-matrix is complete vis-à-vis a particular DCM. The approach developed here relies on the theoretical framework of general DCMs (de la Torre 2011; Henson, Templin & Willse, 2009; Rupp, Templin & Henson, 2010; Von Davier 2005; von Davier 2008).

2 Review of Technical Key Concepts

Assume a knowledge domain can be characterized by K attributes. Then, there are $M = 2^K$ distinct proficiency classes, each of which is defined by a K-dimensional binary attribute profile $\boldsymbol{\alpha}_m = (\alpha_{m1}, \alpha_{m2}, \ldots, \alpha_{mk} \ldots \alpha_{mK})'$, with $m = 1, 2, \ldots, M$. The parameters of a DCM and examinees' attribute profiles are estimated from their observed responses Y_j, $j = 1, 2, \ldots, J$, to the J items in the test. Each individual item itself is associated with a K-dimensional binary vector $\mathbf{q}_j$ called item attribute profile, where $q_{jk} = 1$ if a correct answer requires mastery of the k^{th} attribute, and 0 otherwise. Note that item attribute profiles consisting entirely of zeroes are inadmissible, because they correspond to items that require no skills at all. Hence, given K attributes, there are at most $2^K - 1$ distinct item attribute profiles. The J item attribute profiles of a test constitute its Q-matrix, $\mathbf{Q} = \{q_{jk}\}_{(J \times K)}$, (Tatsuoka 1985) that summarizes the associations between items and attributes.

Recall that the Q-matrix must fulfill the completeness requirement—formally, $\boldsymbol{S}(\boldsymbol{\alpha}) = \boldsymbol{S}(\boldsymbol{\alpha}^*) \Rightarrow \boldsymbol{\alpha} = \boldsymbol{\alpha}^*$, where $\boldsymbol{S}(\boldsymbol{\alpha}) = E(\boldsymbol{Y} \mid \boldsymbol{\alpha})$ is the (conditional) expectation of item response vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_J)'$, given attribute profile $\boldsymbol{\alpha}$. Verbally stated, a Q-matrix is said to be complete if it allows for the identification of all M possible proficiency classes among examinees (Chiu, Douglas & Li, 2009).

DCMs model the functional relation between attribute mastery and the probability of a correct item response. The distinct parameterizations of specific DCMs reflect differences in the underlying theories on how (non-)mastery of attributes affects an examinee's performance. General DCMs express these distinct functional relations in a unified mathematical form and parameterization. Von Davier (2005, 2008) General Diagnostic Model (GDM) is the archetypal general DCM. The item response function (IRF) of presumably the most popular version of von Davier's GDM is formed by the logistic function of the linear combination of all K attribute main effects α_k. Henson, Templin, and Willse (2009) defined the IRF of a general DCM called the Log-Linear Cognitive Diagnosis Model (LCDM) as the logistic function of the linear combination of all K attribute main effects, and all their two-way, three-way, …, K-way interactions. Let

$$h_j = \beta_{j0} + \sum_{k=1}^{K} \beta_{jk} q_{jk} \alpha_{ik} + \sum_{k'=k+1}^{K} \sum_{k=1}^{K-1} \beta_{j(kk')} q_{jk} q_{jk'} \alpha_{ik} \alpha_{ik'} + \cdots + \beta_{j12\ldots K} \prod_{k=1}^{K} q_{jk} \alpha_{ik} \quad (1)$$

where α_{ik} denotes whether examinee i has mastered attribute α_k and q_{jk} indicates whether mastery of attribute α_k is required for item j. The IRF of the LCDM is then

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i) = \frac{\exp(h_j)}{1 + \exp(h_j)} \quad (2)$$

By imposing appropriate constraints on the β-coefficients in h_j, the IRFs of specific DCMs can be expressed as submodels of the LCDM. (The examinee index i is henceforth omitted for brevity if the context permits.)

3 An Analysis of the Conditions of Completeness of the Q-Matrix

Recall that completeness of a given Q-matrix can only be determined in reference to a particular DCM because $\mathbf{Q}$ can be complete for one DCM but incomplete for another. When reparameterizing specific DCMs in terms of the LCDM, the particular composition of h_j allows for the distinction between DCMs with main effects only and DCMs with main effects and interaction effects. A third group consists of DCMs containing only interaction effects. Two special cases are the Deterministic Input Noisy Output "AND" gate (DINA) model (Junker & Sijtsma, 2001; Macready & Dayton, 1977) and the Deterministic Input Noisy Output "OR" gate (DINO) model (Templin & Henson, 2006) that form a fourth group of their own. The examination of conditions of Q-completeness is guided by this categorization of DCMs.

3.1 The DINA Model and the DINO Model

3.1.0.2 The Deterministic Input Noisy Output "AND" Gate Model

Let the set $\mathscr{L}_j = \{k \mid q_{jk} = 1\}$ contain the indices of the non-zero elements in the attribute vector $\mathbf{q}_j$ of item j (i.e., the indices of all attributes α_k required for a correct response to item j). Thus, the IRF of the DINA model reparameterized in terms of the LCDM is

$$P(Y_j = 1 \mid \boldsymbol{\alpha}) = \frac{\exp\left(\beta_{j0} + \beta_{j(\forall k \in \mathscr{L}_j)} \prod_{k \in \mathscr{L}_j} \alpha_k\right)}{1 + \exp\left(\beta_{j0} + \beta_{j(\forall k \in \mathscr{L}_j)} \prod_{k \in \mathscr{L}_j} \alpha_k\right)} \tag{3}$$

subject to $\beta_{j(\forall k \in \mathscr{L}_j)} > 0$. (If $k \in \mathscr{L}_j = \{k \mid q_{jk} = 1\}$, then $q_{jk} = 1$ is always true; hence, q_{jk} has been dropped from the IRF.)

3.1.0.3 The Deterministic Input Noisy Output "OR" Gate Model The DINO model (Templin & Henson, 2006) is a disjunctive CDM—that is, mastery of a subset of the required attributes is a sufficient condition for maximizing the probability of a correct item response. If the DINO model is reparameterized as a general DCM using the logit link, then this condition translates into the constraint that all coefficients—except β_{j0}—in h_j be equal. Only their signs oscillate depending on the order c of the terms in h_j: $(-1)^{c+1}$; hence, for main effects, $c = 1$; for two-way interactions $c = 2$, and so on. Hence, the IRF of the DINO model is

$$P(Y_j = 1 \mid \boldsymbol{\alpha}) = \frac{\exp\left(\beta_{j0} + \beta_{jk}\left(1 - \prod_{l \in \mathscr{L}_j}(1 - \alpha_l)\right)\right)}{1 + \exp\left(\beta_{j0} + \beta_{jk}\left(1 - \prod_{l \in \mathscr{L}_j}(1 - \alpha_l)\right)\right)} \text{ for some } k \in \mathscr{L}_j \tag{4}$$

subject to $\beta_{jk} > 0$.

Chiu and collaborators proved for the DINA model (Chiu, Douglas, & Li, 2009) and the DINO model (Chiu & Köhn, 2015) that $\mathbf{Q}$ is complete if and only if each attribute is represented by at least one single-attribute item—that is, $\mathbf{Q}$ has rows, $\mathbf{e}_1, \ldots, \mathbf{e}_K$, among its J rows, where $\mathbf{e}_k$ is a $1 \times K$ vector, with the k^{th} element, e_k, equal to 1, and all other entries equal to 0. As an illustration for the DINA model, consider the two Q-matrices, $\mathbf{Q}_{1:3}$ and $\mathbf{Q}_{4:6}$, each with $K = 3$ attributes and $J = 3$ items

$$\mathbf{Q}_{1:3} = \begin{pmatrix} 0\ 1\ 1 \\ 1\ 0\ 1 \\ 1\ 1\ 0 \end{pmatrix} \quad \mathbf{Q}_{4:6} = \begin{pmatrix} 1\ 0\ 0 \\ 0\ 1\ 0 \\ 0\ 0\ 1 \end{pmatrix}$$

with the matrix subscripts referring to the item indices $j = 1, \ldots, 6$. $\mathbf{Q}_{1:3}$ is not complete, whereas $\mathbf{Q}_{4:6}$ is complete, as the computation of the expected item-response profiles $\boldsymbol{S}(\boldsymbol{\alpha})$ demonstrates. For the DINA model, the entries in $\boldsymbol{S}(\boldsymbol{\alpha})$ are defined as

Table 1 The DINA model: Expected item responses $S_j(\boldsymbol{\alpha})$ for distinct proficiency classes $\boldsymbol{\alpha}_m$, given the Q-matrices $\mathbf{Q}_{1:3}$ and $\mathbf{Q}_{4:6}$

	$\mathbf{Q}_{1:3}$			$\mathbf{Q}_{4:6}$		
$\boldsymbol{\alpha}$	$\mathbf{q}_1 = (011)$	$\mathbf{q}_2 = (101)$	$\mathbf{q}_3 = (110)$	$\mathbf{q}_4 = (100)$	$\mathbf{q}_5 = (010)$	$\mathbf{q}_6 = (001)$
	$S_1(\boldsymbol{\alpha})$	$S_2(\boldsymbol{\alpha})$	$S_3(\boldsymbol{\alpha})$	$S_4(\boldsymbol{\alpha})$	$S_5(\boldsymbol{\alpha})$	$S_6(\boldsymbol{\alpha})$
(000)	β_{10}	β_{20}	β_{30}	β_{40}	β_{50}	β_{60}
(100)	β_{10}	β_{20}	β_{30}	$\beta_{40} + \beta_{41}$	β_{50}	β_{60}
(010)	β_{10}	β_{20}	β_{30}	β_{40}	$\beta_{50} + \beta_{52}$	β_{60}
(001)	β_{10}	β_{20}	β_{30}	β_{40}	β_{50}	$\beta_{60} + \beta_{63}$
(110)	β_{10}	β_{20}	$\beta_{30} + \beta_{3(12)}$	$\beta_{40} + \beta_{41}$	$\beta_{50} + \beta_{52}$	β_{60}
(101)	β_{10}	$\beta_{20} + \beta_{2(13)}$	β_{30}	$\beta_{40} + \beta_{41}$	β_{50}	$\beta_{60} + \beta_{63}$
(011)	$\beta_{10} + \beta_{1(23)}$	β_{20}	β_{30}	β_{40}	$\beta_{50} + \beta_{52}$	$\beta_{60} + \beta_{63}$
(111)	$\beta_{10} + \beta_{1(23)}$	$\beta_{20} + \beta_{2(13)}$	$\beta_{30} + \beta_{3(12)}$	$\beta_{40} + \beta_{41}$	$\beta_{50} + \beta_{52}$	$\beta_{60} + \beta_{63}$

$$S_j(\boldsymbol{\alpha}) = E(Y_j \mid \boldsymbol{\alpha}) = P(Y_j = 1 \mid \boldsymbol{\alpha}) = \frac{\exp\left(\beta_{j0} + \beta_{j(\forall k \in \mathscr{L}_j)} \prod_{k \in \mathscr{L}_j} \alpha_k\right)}{1 + \exp\left(\beta_{j0} + \beta_{j(\forall k \in \mathscr{L}_j)} \prod_{k \in \mathscr{L}_j} \alpha_k\right)} \tag{5}$$

Table 1 only reports the coefficients that are retained in $S_j(\boldsymbol{\alpha})$, but not the expression of the entire logistic function. Clearly, $\mathbf{Q}_{1:3}$ is not complete because, for example, $\boldsymbol{\alpha}_1 = (000)' \neq \boldsymbol{\alpha}_2 = (100)'$, but $\boldsymbol{S}(\boldsymbol{\alpha}_1) = \boldsymbol{S}(\boldsymbol{\alpha}_2) = (\frac{e^{\beta_{10}}}{1+e^{\beta_{10}}}, \frac{e^{\beta_{20}}}{1+e^{\beta_{20}}}, \frac{e^{\beta_{30}}}{1+e^{\beta_{30}}})$. Thus, $\mathbf{Q}_{1:3}$ does not allow to distinguish between all $\boldsymbol{\alpha}$ (i.e., all the $M = 2^K$ proficiency classes). However, if items 4–6 of $\mathbf{Q}_{4:6}$ are included, then $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^* \Rightarrow \boldsymbol{S}(\boldsymbol{\alpha}) \neq \boldsymbol{S}(\boldsymbol{\alpha}^*)$ because these three single-attribute items have $\mathbf{q}_j = \mathbf{e}_k$; hence, the term $\beta_{j(\forall k \in \mathscr{L}_j)} \prod_{k \in \mathscr{L}_j} \alpha_k$ is reduced to an attribute "main effect"—β_{jk}—that then allows for discriminating between $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\boldsymbol{\alpha}_3$, and $\boldsymbol{\alpha}_4$.

In summary, for the DINA model and the DINO model, the inclusion of all K single-attribute items in the Q-matrix is a necessary condition for its completeness. For other CDMs, however, this is a sufficient, but not a necessary condition—that is, alternative compositions of the Q-matrix that do not include all single-attribute items also guarantee completeness, as is demonstrated in the next section.

3.2 DCMs With Main Effects Only

As an example for a DCM with main effects only, consider the GDM that has IRF and expected item response $S_j(\boldsymbol{\alpha})$

$$P(Y_j = 1 \mid \boldsymbol{\alpha}) = \frac{\exp\left(\beta_{j0} + \sum_{k=1}^{K} \beta_{jk} q_{jk} \alpha_k\right)}{1 + \exp\left(\beta_{j0} + \sum_{k=1}^{K} \beta_{jk} q_{jk} \alpha_k\right)} = S_j(\boldsymbol{\alpha}) \tag{6}$$

Table 2 The GDM: Expected item responses $S_j(\boldsymbol{\alpha})$ for distinct proficiency classes $\boldsymbol{\alpha}_m$, given the Q-matrix $\mathbf{Q}_{1:3}$

	$\mathbf{Q}_{1:3}$		
	$\mathbf{q}_1 = (011)$	$\mathbf{q}_2 = (101)$	$\mathbf{q}_3 = (110)$
$\boldsymbol{\alpha}$	$S_1(\boldsymbol{\alpha})$	$S_2(\boldsymbol{\alpha})$	$S_3(\boldsymbol{\alpha})$
(000)	β_{10}	β_{20}	β_{30}
(100)	β_{10}	$\beta_{20} + \beta_{21}$	$\beta_{30} + \beta_{31}$
(010)	$\beta_{10} + \beta_{12}$	β_{20}	$\beta_{30} + \beta_{32}$
(001)	$\beta_{10} + \beta_{13}$	$\beta_{20} + \beta_{23}$	β_{30}
(110)	$\beta_{10} + \beta_{12}$	$\beta_{20} + \beta_{21}$	$\beta_{30} + \beta_{31} + \beta_{32}$
(101)	$\beta_{10} + \beta_{13}$	$\beta_{20} + \beta_{21} + \beta_{23}$	$\beta_{30} + \beta_{31}$
(011)	$\beta_{10} + \beta_{12} + \beta_{13}$	$\beta_{20} + \beta_{23}$	$\beta_{30} + \beta_{32}$
(111)	$\beta_{10} + \beta_{12} + \beta_{13}$	$\beta_{20} + \beta_{21} + \beta_{23}$	$\beta_{30} + \beta_{31} + \beta_{32}$

For the GDM, $\mathbf{Q}_{4:6}$ is guaranteed to be complete due to the sufficiency condition. However, $\mathbf{Q}_{1:3}$ is also complete for the GDM despite the removal of all interaction effects $\beta_{j(kk')}$ from the model—that is, $\mathbf{S}(\boldsymbol{\alpha}) = \mathbf{S}(\boldsymbol{\alpha}^*) \Rightarrow \boldsymbol{\alpha} = \boldsymbol{\alpha}^*$ still holds (see Table 2):

3.3 DCMs with Main Effects and Interaction Effects

Take the (saturated) LCDM as an example for a model containing all main effects and all interaction effects. For $K = 3$ attributes, the IRF is

$$P(Y_j{=}1 \mid \boldsymbol{\alpha}) = \frac{\exp\left(\beta_{j0} + \sum_{k=1}^{3} \beta_{jk} q_{jk} \alpha_k + \sum_{k'=k+1}^{3} \sum_{k=1}^{2} \beta_{j(kk')} q_{jk} q_{jk'} \alpha_k \alpha_{k'} + \beta_{j(123)} \prod_{k=1}^{3} q_{jk} \alpha_k\right)}{1 + \exp\left(\beta_{j0} + \sum_{k=1}^{3} \beta_{jk} q_{jk} \alpha_k + \sum_{k'=k+1}^{3} \sum_{k=1}^{2} \beta_{j(kk')} q_{jk} q_{jk'} \alpha_k \alpha_{k'} + \beta_{j(123)} \prod_{k=1}^{3} q_{jk} \alpha_k\right)} \quad (7)$$

Note that the expression of the expected response $S_j(\boldsymbol{\alpha})$ is equal to the IRF of item j. For the saturated LCDM, $\mathbf{Q}_{4:6}$ is complete due to the sufficiency condition that Q-matrices containing all K single-attribute items are complete. $\mathbf{Q}_{1:3}$, on the other hand, does not contain any single-attribute item, but is also complete for the saturated LCDM, as the calculation of the $\boldsymbol{S}(\boldsymbol{\alpha})$ reported in Table 3 shows.

3.4 DCMs With No Main Effects, But Only Interaction Effects

What are the consequences if all main effects are removed from, say the saturated LCDM? As an example, consider again the case of $K = 3$; Eq. (8) is the IRF of the no-main-effects model.

Table 3 The saturated LCDM: Expected item responses $S_j(\boldsymbol{\alpha})$ for distinct proficiency classes $\boldsymbol{\alpha}_m$, given the Q-matrix $\mathbf{Q}_{1:3}$

	$\mathbf{Q}_{1:3}$		
	$\mathbf{q}_1 = (011)$	$\mathbf{q}_2 = (101)$	$\mathbf{q}_3 = (110)$
$\boldsymbol{\alpha}$	$S_1(\boldsymbol{\alpha})$	$S_2(\boldsymbol{\alpha})$	$S_3(\boldsymbol{\alpha})$
(000)	β_{10}	β_{20}	β_{30}
(100)	β_{10}	$\beta_{20} + \beta_{21}$	$\beta_{30} + \beta_{31}$
(010)	$\beta_{10} + \beta_{12}$	β_{20}	$\beta_{30} + \beta_{32}$
(001)	$\beta_{10} + \beta_{13}$	$\beta_{20} + \beta_{23}$	β_{30}
(110)	$\beta_{10} + \beta_{12}$	$\beta_{20} + \beta_{21}$	$\beta_{30} + \beta_{31} + \beta_{32} + \beta_{3(12)}$
(101)	$\beta_{10} + \beta_{13}$	$\beta_{20} + \beta_{21} + \beta_{23} + \beta_{2(13)}$	$\beta_{30} + \beta_{31}$
(011)	$\beta_{10} + \beta_{12} + \beta_{13} + \beta_{1(23)}$	$\beta_{20} + \beta_{23}$	$\beta_{30} + \beta_{32}$
(111)	$\beta_{10} + \beta_{12} + \beta_{13} + \beta_{1(23)}$	$\beta_{20} + \beta_{21} + \beta_{23} + \beta_{2(13)}$	$\beta_{30} + \beta_{31} + \beta_{32} + \beta_{3(12)}$

Table 4 No-main-effects model: Expected item responses $S_j(\boldsymbol{\alpha})$ for distinct proficiency classes $\boldsymbol{\alpha}_m$, given the incomplete Q-matrices $\mathbf{Q}_{1:3}$ and $\mathbf{Q}_{4:6}$

	$\mathbf{Q}_{1:3}$			$\mathbf{Q}_{4:6}$		
	$\mathbf{q}_1 = (011)$	$\mathbf{q}_2 = (101)$	$\mathbf{q}_3 = (110)$	$\mathbf{q}_4 = (100)$	$\mathbf{q}_5 = (010)$	$\mathbf{q}_6 = (001)$
$\boldsymbol{\alpha}$	$S_1(\boldsymbol{\alpha})$	$S_2(\boldsymbol{\alpha})$	$S_3(\boldsymbol{\alpha})$	$S_4(\boldsymbol{\alpha})$	$S_5(\boldsymbol{\alpha})$	$S_6(\boldsymbol{\alpha})$
(000)	β_{10}	β_{20}	β_{30}	β_{40}	β_{50}	β_{60}
(100)	β_{10}	β_{20}	β_{30}	β_{40}	β_{50}	β_{60}
(010)	β_{10}	β_{20}	β_{30}	β_{40}	β_{50}	β_{60}
(001)	β_{10}	β_{20}	β_{30}	β_{40}	β_{50}	β_{60}
(110)	β_{10}	β_{20}	$\beta_{30} + \beta_{3(12)}$	β_{40}	β_{50}	β_{60}
(101)	β_{10}	$\beta_{20} + \beta_{2(13)}$	β_{30}	β_{40}	β_{50}	β_{60}
(011)	$\beta_{10} + \beta_{1(23)}$	β_{20}	β_{30}	β_{40}	β_{50}	β_{60}
(111)	$\beta_{10} + \beta_{1(23)}$	$\beta_{20} + \beta_{2(13)}$	$\beta_{30} + \beta_{3(12)}$	β_{40}	β_{50}	β_{60}

$$P(Y_j{=}1 \mid \boldsymbol{\alpha}){=}\frac{\exp\left(\beta_{j0}{+}\sum_{k'=k+1}^{3}\sum_{k=1}^{2}\beta_{j(kk')}q_{jk}q_{jk'}\alpha_k\alpha_{k'}{+}\beta_{j(123)}\prod_{k=1}^{3}q_{jk}\alpha_k\right)}{1{+}\exp\left(\beta_{j0}{+}\sum_{k'=k+1}^{3}\sum_{k=1}^{2}\beta_{j(kk')}q_{jk}q_{jk'}\alpha_k\alpha_{k'}{+}\beta_{j(123)}\prod_{k=1}^{3}q_{jk}\alpha_k\right)} \tag{8}$$

Then, as the inspection of the $\boldsymbol{S}(\boldsymbol{\alpha})$ reported in Table 4 immediately shows, matrix $\mathbf{Q}_{1:3}$ is no longer complete because some $\boldsymbol{S}(\boldsymbol{\alpha}) = \boldsymbol{S}(\boldsymbol{\alpha}^*)$ despite $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$. Thus, four of the proficiency classes are not identifiable. Note that, different from the DINA model, using $\mathbf{Q}_{4:6}$ as Q-matrix instead of $\mathbf{Q}_{1:3}$ does not resolve the completeness issue but rather seems to worsen it because then, none of the proficiency classes is identifiable (see Table 4).

Table 5 Main-effects-only model: Expected item responses $S_j(\boldsymbol{\alpha})$ for proficiency classes $\boldsymbol{\alpha} = (001)$ and $\boldsymbol{\alpha} = (110)$, given the Q-matrix $\mathbf{Q}$

	$\mathbf{Q}$		
	$\mathbf{q}_1 = (101)$	$\mathbf{q}_2 = (011)$	$\mathbf{q}_3 = (111)$
$\boldsymbol{\alpha}$	$S_1(\boldsymbol{\alpha})$	$S_2(\boldsymbol{\alpha})$	$S_3(\boldsymbol{\alpha})$
(001)	$\beta_{10} + \beta_{13}$	$\beta_{20} + \beta_{23}$	$\beta_{30} + \beta_{33}$
(110)	$\beta_{10} + \beta_{11}$	$\beta_{20} + \beta_{22}$	$\beta_{30} + \beta_{31} + \beta_{32}$

4 Rules of Q-Completeness

In light of the last result, it comes as no surprise that models containing no main effects, but only interaction effects—at least to our knowledge—have never been proposed in the literature: These models cannot discriminate between the M proficiency classes. Said differently, for models without the k^{th} main effect, any Q-matrix is incomplete.

The DINA model and the DINO model form a category of their own: A Q-matrix to be used with either of the two models is complete if and only if it contains among its J items all K single-attribute items having item attribute vectors $\mathbf{q}_j = \mathbf{e}_k$, where $\mathbf{e}_k$ was defined earlier as a unit vector with all elements equal 0 except the k^{th} entry (for proofs of this claim, consult Chiu, Douglas, & Li, 2009; Chiu & Kohn, 2015).

For DCMs containing only main effects, consider two K-dimensional attribute profiles $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$. Then there exists at least one k such that $\alpha_k = 1$ and $\alpha_k^* = 0$. In addition, assume that q_{jk} in $\mathbf{Q}$ is 1 for some j. Thus, for models that contain only main effects, a $J \times K$ matrix $\mathbf{Q}$ is complete if and only if it contains K linearly independent q-vectors and $\sum_{k'=1,k'\neq k}^{K} \beta_{jk'} q_{jk'} (\alpha_{k'} - \alpha_{k'}^*) \neq \beta_{jk}$ for some k. As an example, consider

$$\mathbf{Q} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

that consists of three linearly independent q-vectors. But the constraint $\sum_{k'=1,k'\neq k}^{K} \beta_{jk'} q_{jk'} (\alpha_{k'} - \alpha_{k'}^*) \neq \beta_{jk}$ is possibly violated, as the inspection of the $\boldsymbol{S}(\boldsymbol{\alpha})$ reported in Table 5 implies: If $\beta_{13} = \beta_{11}$, $\beta_{23} = \beta_{22}$, and $\beta_{33} = \beta_{31} + \beta_{32}$, then the two proficiency classes with attribute profiles (001) and (110) cannot be distinguished. However, this particular constellation is pretty rare; it can only occur if the expected responses for distinct $\boldsymbol{\alpha}$ are not nested within each other.

For DCMs containing main effects and interaction effects, consider two attribute profiles $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$. Then there exists at least one k such that $\alpha_k = 1$ and $\alpha_k^* = 0$. In addition, assume that q_{jk} in $\mathbf{Q}$ is 1 for some j. Hence, for models that contain main effects and interaction terms, a $J \times K$ matrix $\mathbf{Q}$ is complete if and only if it contains K linearly independent q-vectors and $\sum_{k'=1,k'\neq k}^{K} \beta_{jk'} q_{jk'} (\alpha_{k'} - \alpha_{k'}^*) +$

Table 6 Main-and-interaction-effects model: Expected item responses $S_j(\boldsymbol{\alpha})$ for proficiency classes $\boldsymbol{\alpha} = (001)$ and $\boldsymbol{\alpha} = (110)$, given the Q-matrix $\mathbf{Q}$

	$\mathbf{Q}$		
	$\mathbf{q}_1 = (101)$	$\mathbf{q}_2 = (011)$	$\mathbf{q}_3 = (111)$
$\boldsymbol{\alpha}$	$S_1(\boldsymbol{\alpha})$	$S_2(\boldsymbol{\alpha})$	$S_3(\boldsymbol{\alpha})$
(001)	$\beta_{10} + \beta_{13}$	$\beta_{20} + \beta_{23}$	$\beta_{30} + \beta_{33}$
(110)	$\beta_{10} + \beta_{11}$	$\beta_{20} + \beta_{22}$	$\beta_{30} + \beta_{31} + \beta_{32} + \beta_{3(12)}$

$\cdots + \beta_{j(12\ldots K)} \prod_{k=1}^{K} q_{jk} \left(\prod_{k=1}^{K} \alpha_k - \prod_{k=1}^{K} \alpha_k^* \right) \neq -\beta_{jk}$ for some k. Consider again $\mathbf{Q}$ used in the previous example as an illustration. Unless the constraints $\beta_{13} \neq \beta_{11}$, $\beta_{23} \neq \beta_{22}$, and $\beta_{33} \neq \beta_{31} + \beta_{32} + \beta_{3(12)}$ are in effect, the two proficiency classes with attribute profiles (001) and (110) cannot be distinguished (see Table 6).

As a concluding remark, the answer to the question whether the rules for determining completeness of the Q-matrix are also applicable if the attributes have a hierarchical structure awaits further research. At present, it is not clear to what extent the varying complexity of different attribute hierarchies might affect the usefulness of the criteria for Q-completeness described earlier—in not mentioning the further complication that multiple hierarchies possibly underlie the structural relation among attributes.

References

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*, 633–665.

Chiu, C.-Y., & Köhn, H.-F. (2015). Consistency of cluster analysis for cognitive diagnosis: The DINO model and the DINA model revisited. *Applied Psychological Measurement, 39*, 465–479.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199.

DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics*. Psychometrics (Vol. 26, pp. 979–1030). Amsterdam: Elsevier.

Haberman, S. J., & von Davier, M. (2007). Some notes on models for cognitively based skill diagnosis. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics*. Psychometrics (Vol. 26, pp. 1031–1038). Amsterdam: Elsevier.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191–210.

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

Leighton, J., & Gierl, M. (2007) *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge: Cambridge University Press.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 33*, 379–416.

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

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconception in the pattern classification approach. *Journal of Educational Statistics, 12*, 55–73.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305.

von Davier, M. (2005, September). *A general diagnostic model applied to language testing data* (Research Rep. No. RR-05-16). Princeton: Educational Testing Service.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–301.

Application Study on Online Multistage Intelligent Adaptive Testing for Cognitive Diagnosis

Fen Luo, Shuliang Ding, Xiaoqing Wang, and Jianhua Xiong

Abstract "On-the-fly assembled multistage adaptive Testing (OMST)" provides some unique advantages for both Computerized Adaptive Testing (CAT) and Multistage Testing (MST). In OMST, not one but multiple items are assembled on the fly into one unit in each stage. We apply the idea of OMST to Cognitive Diagnosis CAT (CD-CAT), name it as Online Multistage Intelligent Adaptive Testing (OMIAT), which aims to accurately estimate both examinees' latent ability level and their knowledge state (KS) simultaneously. A simulation study was conducted to five different item selection methods in CD-CAT: OMIAT method, Shannon Entropy (SHE) method, Aggregate standardized information (ASI) method, Maximum Fisher Information (MFI) method, and random method. The result shows that: (1) both the OMIAT and the ASI methods can not only measure the ability level with precision, but also classify the examinee's KS with accuracy. In most cases, the OMIAT method is superior to the ASI method in terms of the evaluation criteria, especially when the number of attributes, which is required to respond correctly to the item, is small (<=2). (2) The pattern classification correct rate of the SHE method is always the highest and that of the OMIAT method is always second, but the item exposure rate and the time consumption of the OMIAT method is far superior to those of the SHE method.

Keywords Cognitive diagnosis • Adaptive testing • Item Response Theory • Online multistage adaptive testing • Item selection method

1 Introduction

During the long-term process of using Computerized Adaptive Testing (CAT), people have discovered some of its defects. For example, in 2000, Educational Testing Service (ETS) found that the Graduate Record Examination (GRE) CAT system did not produce reliable scores for a few thousand examinees (Carlson 2000;

F. Luo (✉) • S. Ding (✉) • X. Wang • J. Xiong
School of Computer and Information Engineering, Jiangxi Normal University,
99 Ziyang Ave., 330022 Nanchang, Jiangxi, China
e-mail: luofen312@163.com; ding06026@163.com; wxqfree@163.com; pansy1212@sina.com

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer
Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_20

Chang 2004); CAT did not allow examinees to skip items or revisit completed items and there was a lack of control over the non-statistical properties of the tests forms before administration (Hendrickson 2007). To offset some of its disadvantages, the multi-stage adaptive test (MST) was proposed. In MST, a test is comprised of several different stages with each stage having a certain number of modules, which include several items in each module, anchored at varied difficulty levels. Only one module of each stage will be selected in the real exam. The whole test structure must be prepared before the administration. Recently, the On-the-fly MST (OMST) is addressed that it combines the advantages of CAT and MST and offsets their limitations (Chang 2015; Zheng & Chang 2015). Like MST, OMST is administered in stages and only adapts between stages. But different from MST, where the modules to be administrated in each stage are selected from several pre-assembled modules of that stage; the modules to be administrated in each stage in OMST are assembled on the fly.

CAT focuses on providing better ability estimation with a shorter test. Cognitive diagnosis models (CDMs) have been developed to detect mastery and non-mastery of attributes or skills. Cognitive diagnosis CAT (CD-CAT) can achieve the same performance on knowledge state (KS) estimate as CDMs with fewer items.

Both the implementation of CD-CAT and the item selection methods depend on CDMs. Many CDMs have been proposed (Rupp, Templin & Henson 2010), and the Deterministic Inputs, Noisy—and- gate (DINA) (Haertel 1989; Junker & Sijtsma 2001) is easy to explain and operate, and widely used in researches of Cognitive Diagnosis and CD-CAT.

Shannon Entropy (SHE) (Xu, Chang & Douglas 2003) and Kullback–Leibler (KL) (Cover & Thomas 1991) information are famous indices in CD-CAT. There are several variations selection methods on KL, for instance, the Posterior-Weighted KL (PWKL) index (Cheng 2009), Aggregate standardized information (ASI) method (Wang, Zheng & Chang 2014) and so on.

CAT focuses on measuring latent ability level precisely and CD-CAT focuses on classifying the student according to KS accurately. McGlohen and Chang (2008), Cheng and Chang (2007), Wang, Chang, and Douglas (2012), Wang, Zheng, and Chang (2014) solved the dual-objective, namely by not only estimating latent ability level efficiently, but also classifying the student's KS accurately.

Like in CAT, items are administrated one by one in CD-CAT. In MST, there are time-consuming processes including the test design, assembly methods, and routing rules. In this study, we combined CD and OMST to build a new test design method named Online Multistage Intelligent Adaptive Testing (OMIAT), which we examine in a simulation study in comparison with other well-known methods. OMIAT has the following characteristics: (1) Its goal is to accurately estimate examinees' latent ability levels and KS simultaneously, (2) Routing rules and items assembly are automatically planned.

2 OMIAT

Let θ be the unidimensional continuous latent ability to be measured and $\alpha = (\alpha_1, \cdots, \alpha_K)$ be the *K-dimension* KS to be measured (K is the number of attributes) in the test. The value of the vector's kth element is 1 if the examinee has mastered the kth attribute; otherwise, it is 0.

2.1 Important Concepts

1. Adjacency matrix and Reachability matrix

   The adjacency matrix (denoted by A) represents the direct hierarchical relation among the attributes. For example, $a_{ij} = 1$ means the ith attribute is the immediate prerequisites to the jth attribute.

   The reachability matrix (denoted by R) represents a direct or indirect relationship among the attributes, $r_{ij} = 1$ means the ith attribute is the direct or indirect prerequisite to the jth attribute. For the independent attribute hierarchy, the adjacency matrix is a matrix with all elements being zero, and the reachability matrix is an identity matrix.
2. Q-matrix theory

   In Q-matrix theory (Tatsuoka 1995, 2009), which plays a pivotal role in CDMs, the Q-matrix is a matrix that relates the items to the attributes. Let Q be a $K \times J$ matrix, and each column of the Q-matrix represents a kind of a potential item type (K is the number of attributes, J is the number of potential items). Q matrix's element q_{kj} is 1 if the kth attribute is required to respond correctly to the jth potential item, otherwise it is 0. The columns of a Q-matrix are a subset of all possible potential item types.

   Q-matrix theory first tries to build the equivalence relationship between examinee's KS and expected response pattern (ERP), then map the observed response pattern (ORP) to the closest ERP through some classification methods, so we can finally find the KS behind the ORP. But Tatsuoka (1995, 1995, 2009) didn't seem to attain this goal.

   The complement of Q-matrix theory (Ding, Luo, Cai, Lin & Wang 2008; Ding, Yang & Wang 2010) corrects its imperfections, which includes obtaining a reachability matrix from adjacency matrix, finding a more convenient way to construct a reduced Q matrix and calculate ERPs, and discovering the fact that any column in the Q-matrix can be represented by the combination of the columns of the reachability matrix, so the reachability matrix is a very important special Q-matrix.
3. Lattice theory

   In mathematics, a lattice is a special partially ordered set which contains a unique supremum (also called a least upper bound or join) and a unique infimum (also called a greatest lower bound or meet). The intersection and union

operations on the set of KSs can produce a lattice in which the supremum is the union of all KS vectors and the infimum is the intersection of all KS.

4. Bijective mapping

   A bijective mapping or one-to-one correspondence is a function between the elements of two sets (say X and Y), where every element in the set X is paired with exactly one element in the set Y, and vice versa, every element in the set Y is paired with exactly one element in the set X. The mapping from the set of ERPs to the set of the KSs is a bijective mapping, which means that there are as many ERPs as KSs.

5. MAP

   In Bayesian statistics, maximum a posteriori (MAP) probability estimate is a mode of the posterior distribution. The MAP estimation can be used to obtain a point estimate of an unobserved quantity on the basis of empirical data.

6. HO-DINA

   The higher-order latent trait models (de la Torre & Douglas 2004) combine the Item Response Theory (IRT) model and diagnostic model by assuming conditional independence of response Y given α and also by assuming that the components of α are independent condition on θ. If the examinee's response follows the DINA model given α, then the higher-order latent trait model is called the higher-order DINA model (HO-DINA). de la Torre and Douglas (2004) demonstrated that when fitted with the same data, the value of θ obtained by the HO-DINA model will correlate highly with the value of θ obtained by the two parameters (2PL) IRT model. Therefore, by generating data from the HO-DINA model, we can have two sets of parameters, one from the 2PL model, including discrimination parameter a, difficulty parameter b, and latent ability level θ, which are ready for the unidimensional IRT, and the other set from the DINA model, including slipping parameter, guessing parameter and α, which are requested by cognitive diagnosis (Wang, Chang & Douglas 2012).

2.2 Design of OMIAT

The object of the OMIAT method is not only to yield higher classification precision for α, but also to achieve more accurate estimation for θ. OMIAT is also administered in stages and adapts between stages like OMST. In OMIAT, the new set of items is assembled according to a provisional KS α, which is estimated based on responses of the examinee's finished items up to now. According to the complement of Q-matrix theory by Ding et al. (2010), if the reachability matrix R is a submatrix of the test Q-matrix, it can be guaranteed that we can attain a bijective mapping from the set of ERPs to the set of the KSs, so in the first stage, for each column (i.e. a potential item) of the reachability matrix R, we select one corresponding item into the stage's module. We use set T_i to record all potential items administered in all previous i stages, a provisional α_i vector estimated by MAP can be computed based

on all responses in 1, 2, ..., i stages, and a new set of potential items T_{i+1} can be assembled as follows:

Let $L = \alpha_i \cap T_i$ (means each element of L comes from the intersection operation between α_i and each column of T_i), set $U = \alpha_i \cup T_i$ (means each element of U comes from union operation between α_i and each column of T_i), then $T_{i+1} = (L \cup U)\text{-}T_i$. This process continues until the test is terminated.

For example, we assume that all attributes are independent and the number of attributes is fixed to $K = 5$, so there are 2^K possible KSs and $2^K\text{-}1 = 31$ potential item types except zero-vector.

1. The first stage: $T_1 = \{(1,0,0,0,0), (0,1,0,0,0), (0,0,1,0,0), (0,0,0,1,0), (0,0,0,0,1)\}$,
2. Normally, there are many items corresponding to a given $t_p \in T_i$ in the item pool, among these items, the item that minimizes the expected Shannon entropy of the posterior distributed of α is selected. Note that the expected Shannon entropy is computed based only on those items that meet the t_p, not all items in the item pool.
3. After items of the ith stage are administered to the examinee, the α_i is estimated by MAP and θ_i is estimated by Expected a Posteriori (EAP). The estimated α_i is assumed to (1, 1, 0, 0, 0). If the posterior probability of α_i exceeds 0.9, go to step (5), otherwise go to step (4).
4. Compute T_{i+1}: For example, if $i = 1$, then $L = \alpha_i \cap T_i = \{(1, 0, 0, 0, 0), (0, 1, 0, 0, 0), (0, 0, 0, 0, 0)\}$, $U = \alpha_i \cup T_i = \{(1, 1, 0, 0, 0), (1, 1, 1, 0, 0), (1, 1, 0, 1, 0), (1, 1, 0, 0, 1)\}$, $T_{i+1} = L \cup U\text{-}\ T_i = \{(1, 1, 0, 0, 0), (1, 1, 1, 0, 0), (1, 1, 0, 1, 0), (1, 1, 0, 0, 1)\}$. If T_{i+1} isn't ϕ and α_i isn't (0, 0, 0, 0, 0), then repeat step (2) to (4), otherwise, go to step (5).
5. Select one item from all items which haven't been administered yet using the SHE algorithm.
6. If termination condition is met, stop and exit, otherwise:
   (a) If the maximum posterior probability of α_i exceeds 0.9, go to step (7);
   (b) Otherwise, go to step (5).
7. Select one item using the maximize Fisher item information (MFI) (Lord 1980) at the examinee's current estimated trait level, go to step (6).

3 Simulation Study

The simulation study aimed to investigate the efficiency of the OMIAT compared with SHE, ASI, MFI and Random (RND) selection methods for four item pools with different structures. Pattern correct rate, mean absolute bias, average exposure rate and time consuming were calculated to compare the efficiency of five item selection indices.

$$Q = \left(\begin{array}{ccccc|cccccccccc|cccccccccc|ccccc|c} 1&0&0&0&0&1&1&1&1&0&0&0&0&0&0&1&1&1&1&1&1&0&0&0&0&1&1&1&1&1&0&1\\ 0&1&0&0&0&1&0&0&0&1&1&1&0&0&0&1&1&1&0&0&0&1&1&1&0&0&0&1&1&1&1&1\\ 0&0&1&0&0&0&1&0&0&1&0&0&1&1&0&1&0&0&1&1&0&1&1&0&1&1&0&1&1&1&1&1\\ 0&0&0&1&0&0&0&1&0&0&1&0&1&0&1&0&1&0&1&0&1&1&0&1&1&1&1&0&1&1&0&1&1\\ 0&0&0&0&1&0&0&0&1&0&0&1&0&1&1&0&0&1&0&1&1&0&1&1&1&1&1&1&1&1&0&1&1 \end{array}\right)$$

Fig. 1 Q-matrix

3.1 Experiment Settings

Suppose that the attributes are mutually independent and the number of attributes is $K = 5$, which is a medium number that is often considered in the literature (Wang 2013). The number of all potential items is $2^K - 1 = 31$ as seen in Fig. 1.

Note that a rule of thumb is that the pool should contain at least 12 times as many items as the test length (Stocking 1994). Test length was fixed to 25, and the size of the item pool was fixed to 300. Parameters slipping and guessing of the DINA model were simulated from U (0.05, 0.25) distribution (Hsu, Wang & Chen 2013). We adopt the same parameters settings as the ASI method for the 2PL model parameters (Wang et al. 2014). HO-DINA parameters slope and intercept were chosen such that the result correlations among the attributes were between 0.45 and 0.65 (Segall 1996). A 3000-by-300 complete response matrix was generated based on the HO-DINA model, and it was retrofitted with the 2PL model using the EM algorithm. The item type was defined so that all the items had the same attribute vector, that is to say, they shared the same column of the Q-matrix.

Item bank generation: generate items based on the Q-matrix (see Fig. 1). A 300-item pool was generated with a 300-by-5 Q-matrix. Four item pools were simulated and 1000 examinees were generated for each item pool; each examinee's true KS vector was selected from 2^K α vectors randomly as follows.

1. Study 1: The item pool includes 31 types of potential items, with each potential item type measuring one or five attributes was repeated 15 times, and each potential item type measuring two, three or four attributes was repeated six times. The repeated times were chosen such that the number of items measuring each attribute was as balanced as possible.
2. Study 2: item pool includes 25 types of potential items, with each potential item type measuring one attributes was repeated 28 times, and each potential item type measuring two or three attributes was repeated eight times.
3. Study 3: item pool includes 15 types of potential items, with each potential item type measuring one attribute was repeated 30 times, and each potential item type measuring two attributes was repeated 15 times.
4. Study 4: item pool includes five types of potential items, with each potential item type only measuring one attribute and was repeated 60 times.

In the OMIAT, SHE, ASI, RND selection methods, an examinee's response to each item in a test was generated from the DINA model. In the MFI selection method, examinee responses to each item in a test were generated from the 2PL model.

3.2 Evaluation Criteria

The CD-CAT administration code was written in Python 2.6 and ran on a computer with processor of 2.67 GHz and 3 GB of internal memory, and running time of the program execution is measured in seconds. Four criteria are presented to evaluate the performance of the five item selection methods:

The correct pattern classification rate (PMR) is used to examine accuracy of classification performance; the means of absolute bias error (ABS) is used to evaluate the latent trait estimation precise; the Chi-square index (χ^2) quantifies the efficiency of the item bank usage; the average test consuming time (Tc) is used to evaluate computation speed. These statistics are defined as follows (Wang et al. 2012):

$$\begin{aligned} PMR &= \frac{1}{N}\sum_{i=1}^{N} I\left\{\alpha_i = \widehat{\alpha}_i\right\}, \\ ABS &= \frac{1}{N}\sum_{i=1}^{N} \left|\widehat{\theta}_i - \theta_i\right|, \\ \chi^2 &= \frac{1}{N}\sum_{i=1}^{N} \left(er_j - \overline{er}_j\right)/\overline{er}_j, \\ \text{and} \quad Tc &= \frac{1}{N}\sum_{i=1}^{N} t_i \end{aligned}$$

where N is the examinee sample size, $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{ik})$ and $\widehat{\alpha}_i = (\widehat{\alpha}_{i1}, \ldots, \widehat{\alpha}_{ik})$ represent the true KS and the estimated KS of examinee i, respectively, and $\widehat{\theta}_i$ is the final EAP estimate for examinee i; θ_i is the corresponding true value from either the 2PL or the HO-DINA; er_j is the exposure rate of item j; L is test length and $\overline{er}_j = L/N$ is the desirable uniform rate for all items; t_i is the time which examinee i spent finishing a test. The average item administration time per examinee was recorded separately for each selection method. For PMR, a higher value is better; for the others criteria, lower is better.

3.3 Results and Conclusions

Five different item selection methods are considered in this simulation study. The MFI method would be considered as a baseline which evaluated the accuracy of latent ability level θ, the RND method is the overall baseline, which is non-adaptive with respect to both latent ability level θ and KS α.

Table 1 Results of OMIAT, SHE, ASI, MFI and RND

		PMR	ABS	χ^2	T_C
Study 1	OMIAT	95.5 %	0.32	18.90	1.30
	SHE	**98.0 %**	0.38	87.40	9.14
	ASI	95.1 %	**0.26**	77.69	12.62
	MFI	20.6 %	**0.26**	102.90	**0.16**
	RND	75.1 %	0.36	**0.35**	0.37
Study 2	OMIAT	95.8 %	0.34	21.38	2.06
	SHE	**97.8 %**	0.34	94.43	9.12
	ASI	95.3 %	**0.30**	87.94	12.59
	MFI	16.6 %	**0.30**	117.36	**0.16**
	RND	77.8 %	0.38	**0.44**	0.37
Study 3	OMIAT	97.6 %	0.35	33.04	2.62
	SHE	**98.2 %**	**0.31**	111.49	8.96
	ASI	94.3 %	0.41	97.80	12.34
	MFI	17.2 %	**0.31**	132.44	**0.16**
	RND	80.2 %	0.40	**0.36**	0.36
Study 4	OMIAT	95.0 %	0.38	76.85	3.26
	SHE	**98.1 %**	**0.24**	168.49	9.17
	ASI	57.1 %	0.51	114.31	12.62
	MFI	0.04 %	**0.24**	119.36	**0.18**
	RND	80.2 %	0.39	**0.42**	0.37

Table 1 presents the results on four item pools with different structures. Several conclusions can be drawn from the results. The SHE, ASI, MFI and RND methods are all one by one item selection methods, and do not care about the numbers and distributions of the potential item when selecting the item, so in Study1–4, both the PMR of SHE, ASI, RND methods and the ABS of MFI method show little differences, except that the PMR of ASI method is very low in Study 4. The SHE method has the highest PMR. The MFI method has the highest accuracy of θ. The RND method has the most evenly item exposure rate. The ASI method tended to select the item which could maximize the weighted sum of KL(θ) and PWKL(α), so it performs good on both PMR of α and ABS of θ.

The OMIAT method generates more accurate estimates of θ and α than the ASI method, and its item exposure rate is far less than the ASI method and is close to the RND method. The OMIAT can ensure the security of the test. It selects item based on T_{i+1} ($T_{i+1} = (L \cup U) - T_i$). It doesn't need to search in the whole item pool, so it can satisfy the high response speed request of CAT. The more potential item types included in the item pool, the smaller χ^2 value and time consuming values are achieved.

4 Discussion

The proposed OMIAT constructs another form of CD-CAT. Tests can generally be divided into two parts: estimate of KS α and estimate of latent ability θ. In the process of KS estimation, several items, which are used to deal with the situation that KS α is either overestimated or underestimated, are selected as a stage. The selection method is related to the cognitive model and the CDM. In the process of latent ability estimate, the items are selected one by one using the MFI method.

The Monte Carlo simulations showed that (1) The OMIAT can gain a high PMR and can estimate the examinee's ability level at the same time; (2) In comparison with the SHE, the OMIAT lost little classification precision, but acquired lower average exposure rates and time consumption. For a real-time system such as CAT, computation efficiency is a very desirable property (Cheng 2009).

OMIAT can satisfy the dual purposes, not only to accurately estimate the overall ability but also to classify a cognitive profile in education setting. It can be applied to CD-CAT, and meet special request such as high-speed, the safety of testing including balance of item exposure rate and the item usage etc.

In our simulation study, each item had two types of parameter sets: one belongs to the unidimensional IRT model, the other belongs to the DINA model. Liu, You, Wang, Ding and Chang (2013) developed and implemented a web-based CD-CAT program for a large-scale English test in China, the Level 2 English Achievement. The parameters of the item pool in the web-based CD-CAT are calibration by the three parameter logistic model and the DINA model. We established the relationship between the unidimensional IRT model and the DINA model by the HO-DINA method. Our approach has its limitations, only when there are high correlations among the attributes, or when the attributes display a linear structure, the associational method can be satisfied.

In the future, it would be of interest to study the following aspects. First, Leighton, Gierl and Hunka (Leighton, Gierl & Hunka 2004) claimed that there existed hierarchical relations among cognitive attributes, these hierarchical relations could reduce the number of potential items, and so OMIAT method had the potential of getting higher performance when these hierarchical relations were considered. Second, in the process of assembling items, the potential item types occurred before were removed. It is possible that slip probability and guess probability have the effect on classification correctness rate. This is possible to overcome by removing a potential item if the occurrences reach some upper threshold. Third, in the design of OMIAT, items are assembled by updated estimated KS. Only when the maximum posterior probability of α_i exceeds 0.9 and pre-determined test length is not met, the MFI method is adopted, which adapt the latent ability. The KS estimation procedure and MFI item selection procedure is separated from each other at the moment, thus, a question of interest is if we can find a way to combine these two procedures more closely?

Acknowledgments This research was supported by the National Natural Science Foundation of China (Grant No. 31360237, 31160203) and Science and technology project in Jiangxi province department of education (Grant No. GJJ13208, GJJ13226, GJJ150356). We would like to thank Prof. Marie Wiberg for many helpful comments and suggestions. Thank Mrs. Li Junlan for correcting the paper.

References

Carlson, S. (2000). ETS finds flaws in the way online GRE rates some students. *Chronicle of Higher Education, 47*(8), A47.

Chang, H. H. (2004). Understanding computerized adaptive testing: From Robbins-Monro to Lord and beyond. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 117–133). Thousand Oaks, CA: Sage.

Chang, H. H. (2015). Psychometrics behind Computerized Adaptive Testing. *Psychometrika, 80*(1), 1–20.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing. *Psychometrika, 74*(4), 619–632.

Cheng, Y., & Chang, H. H. (2007). *Dual information method in cognitive diagnostic computerized adaptive testing*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

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

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353.

Ding, S. L., Luo, F., Cai, Y., Lin, H. J., & Wang, X. B. (2008). Complement to Tatsuoka's Q matrix theory. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 417–423). Tokyo, Japan: Universal Academy Press.

Ding, S. L., Yang, S. Q., & Wang, W. Y. (2010). The importance of reachability matrix in constructing cognitively diagnostic testing. *Journal of Jiangxi Normal University (Natural Science Edition), 34*(5), 490–495.

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.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice, 26*(2), 44–52.

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.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's Rule-Space approach. *Journal of Educational Measurement, 41*(3), 205–237.

Liu, H. Y., You, X. F., Wang, W. Y., Ding, S. L., & Chang, H. H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of Classification, 30*(2), 152–172.

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

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

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*(2), 331–354.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools*. Princeton, NJ: Educational Testing Service (ETS Research Rep. No. 94–5).

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern classification approach. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessments* (pp. 327–359). Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York: Routledge/Taylor & Francis Group.

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement, 73*(6), 1017–1035.

Wang, C., Chang, H. H., & Douglas, J. (2012). Combining CAT with cognitive diagnosis: A weighted item selection approach. *Behavior Research Methods, 44*(1), 95–109.

Wang, C., Zheng, C. J., & Chang, H. H. (2014). An enhanced approach to combine item response theory with cognitive diagnosis in adaptive testing. *Journal of Educational Measurement, 51*(4), 358–380.

Xu, X. L., Chang, H. 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.

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

Dichotomous and Polytomous Q Matrix Theory

Shuliang Ding, Fen Luo, Wenyi Wang, and Jianhua Xiong

Abstract Besides specifying the Q matrix correctly, an extended Q matrix theory should provide some answers to following questions: (a) how to find the universal set of knowledge states (KSs), (b) how to use an algorithm to calculate the set of ideal response patterns (IRPs), (c) how to construct the test Q matrix which establishes one-to-one mapping from the set of KSs to the set of IRPs, and (d) how to mine the hierarchy relation hidden in the Q matrix. In this note, a dichotomous and polytomous Q matrix theory is established and the properties of the polytomous Q matrix are explored to help address the above questions.

Keywords Q matrix theory • Reachability matrix • Polytomous Q matrix

Under non-compensable conditions and for Boolean matrices, Q matrix theory was established by Tatsuoka (1995, 2009) for determining a universal set of unobservable knowledge states (KSs) and for classifying students with their observable item response patterns.

The theory was extended by Ding, et al. (Ding, Luo, & Wang 2012). Besides specifying the Q matrix correctly, the extended Q matrix theory is focused upon answering following questions: (a) how to find the universal set of KSs, (b) how to calculate a set of ideal response patterns (IRPs), (c) how to construct a test Q matrix which establishes one-to-one mapping from the set of KSs to the set of IRPs, and (d) how to mine hierarchy relations hidden in the test Q matrix. Among them, the third question (c), is difficult but very important, because the one-to-one mapping, which is helpful to improve the classification accuracy, can relate each of the unobservable KSs to a unique IRP and, and vice versa. For an independent attribute hierarchical structure, and if the test Q matrix contains an identity matrix, the one-to-one mapping from the set of KSs to the set of IRPs has been discovered (Chiu,

S. Ding (✉) • F. Luo (✉) • W. Wang • J. Xiong
School of Computer and Information Engineering, Jiangxi Normal University, 99 Ziyang Ave., 330022 Nanchang, Jiangxi, China
e-mail: ding06026@163.com; luofen312@163.com; hicosdor@aliyun.com; pansy1212@sina.com

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_21

Douglas, & Li 2009; Samejima 1995; Tatsuoka 1995, 2009). For other attribute hierarchies, according result is presented in Sect. 1.2 as Theorem 1.

In cognitive diagnosis the cognitive model of task performance plays a highly important role in cognitive diagnostic assessment. The model involves some Boolean matrices, such as the adjacency matrix A, the reachability matrix R, and a series of Q matrices. For a given hierarchy, the matrix A represents direct relationships among attributes, and the matrix R represents direct and indirect relationships among attributes. These relationships also form a partial order and can be represented by a Hasse diagram which is an intuitive representation of the cognitive model. Furthermore, according to the extended Q matrix theory, the matrix R is a special test Q matrix, and it plays an important role in designing the cognitive diagnostic test blueprint (Ding, Wang, & Yang 2011; Ding, Yang, & Wang 2010).

The purpose of this study is to elucidate certain properties of some Boolean Q matrices through analyzing the above-mentioned relationships mathematically, and to develop a dichotomous and polytomous Q matrix theory to help address the four questions we asked at the beginning of the paper. The polytomous Q matrix has been found important to characterize large grain sizes of attributes (Chen & de la Torre 2013; Sun, Xin, Zhang, & de la Torre 2013; Zhang 2012) and has a close relation with the Boolean Q matrix.

1 Reachability Matrix R is a Core Element of the Extended Q Matrix Theory

The prerequisite relation among attributes is a partial order (Ding & Luo 2013; Tatsuoka 2009). It is equivalent to the inclusion relationship among the set of row vectors (attribute) in a reachability matrix R. Therefore, the matrix R is a representation of a cognitive model (Ding & Luo 2013).

1.1 Relations Among Some Boolean Matrices

1.1.1 Conversion Between Adjacency Matrix and Reachability Matrix

Warshall's algorithm (Rosen 2003) is an efficient method to obtain the matrix R from the matrix A. Conversely, Clean algorithm (Ding & Luo 2005) can be used to convert R into A. Suppose that there are K attributes, $A_1, A_2, \ldots, A_K$, in a domain of interest. Let I be a $K \times K$ identity matrix, and $A = R - I = (a_{ij})_{K \times K}$. Clean algorithm implements a nested loop: the outer loop variable i is from 1 to K; for each i, the intermediate loop variable j is from 1 to K; for each j, the inner loop variable h is also from 1 to K, and the inner loop body is $a_{ih} := a_{ih} - a_{ih}^* a_{ij}^* a_{jh}$.

1.1.2 Augment Algorithm: From R to Q_p

The Augment algorithm (Ding, Luo, Cai, Lin, & Wang 2008; Yang, Cai, Ding, Lin, & Ding 2008) to obtain Q matrix from R includes following steps:

Step 1. Partition R according to its columns.
Step 2. Let Q = R
Step 3. Let $j = 1$
Step 4. Add r_j to every column from $(j + 1)$th column to the last column of Q, and if a new column is produced, put the new column to the far-right side of Q, i.e., augment Q.
Step 5. j:=$j + 1$, if $j \leq K$ then goto step 4 else stop.

It has been proved from the Augment algorithm that all non-zero columns in Q are combinations of the columns in R (Yang & Ding 2011), and it is found that the combination is often not unique.

The Q matrix is also called the reduced Q matrix, and Tatsuoka (1995, 2009) denoted it by Qr. Here, Q_r is replaced by Q_p because its columns are potential item attribute vectors. A test blueprint, denoted by Q_t, is a submatrix of Q_p for a diagnostic test. The student matrix (Q_s) or the universal set of knowledge states can be obtained by adding a column of zeros to Q_p. In addition, Q_s and inclusion relation defined among the columns of Q_s form a lattice (Ding, Luo, Wang, & Xiong 2015; Yang & Ding 2011). It is interesting to note that a Boolean Lattice can be formed only for the independent hierarchy.

1.1.3 Pairwise Comparisons: Mining Attribute Hierarchy from Q Matrix

Tatsuoka (1995 2009) pointed out that pairwise comparisons between any two row vectors in Q matrix with respect to inclusion relations among attributes will yield the reachability matrix R. More accurately, if R is indeed a submatrix of a Q matrix, the attribute hierarchy may be derived from such pairwise comparisons, and all non-zero KSs can be obtained by using the Augment algorithm based on the Q matrix; otherwise, the derived attribute hierarchy may not be the real one, and consequently the obtained set of non-zero KSs by using the Augment algorithm may not coincide with the real set of KSs (Ding & Luo 2013). New Q matrices may be obtained using the Augment algorithm either directly on a given Q matrix as Q_1, or indirectly on the reachability matrix R corresponding to the given attribute hierarchy as Q_2. Q_1 and Q_2 are identical if R is a submatrix of Q, but more generally, Q_1 is a submatrix of Q_2. Note that sometimes a reachability matrix cannot be derived from a given Q matrix, an extreme example is where there are K attributes, but the number of columns in a test Q matrix is less than K.

1.1.4 Ideal Response Patterns

Unless otherwise stated, the non-compensatory cognitive model and 0–1 scoring are considered below.

The set of IRPs can be obtained if Q_s and Q_t are given. Without loss of generality, it is sufficient to consider an ideal response of some students to some items. For i-th column of Q_s (say, x_i) and j-th column of Q_t (say, y_j), if all of the elements in the difference vector $x_i - y_j$ are non-negative, the ideal response is correct, i.e., the ideal score of student x_i on item y_j is 1; otherwise, it is 0.

1.2 Some Q Matrices with Special Properties

Definition of a sufficient Q-matrix (Tatsuoka 1995, 2009). If the pairwise comparison of attribute vectors in the Q matrix yields the reachability matrix R, then the Q-matrix is said to be sufficient for representing the cognitive model of a domain of interest.

Tatsuoka (1995, 2009) claimed that the sufficient Q matrix will improve construct validity of a cognitive diagnostic test and that the sufficient Q matrix is the core of a knowledge structure. However, the sufficient Q matrix cannot guarantee one-to-one mapping from the set of KSs to the set of IRPs.

Example 1. For an independent attribute hierarchy with three attributes, the elements of Q_1 are as follows: all of the diagonal elements are zero, and the rest elements are one. Obviously, the matrix Q_1 is a sufficient Q matrix, but the KSs (0 0 0), (1 0 0), (0 1 0), (0 0 1) correspond to the same IRP (0 0 0). Q_1 maps all of the KSs to only 5 different IRPs. Under the same condition as above, there is a Q matrix (say, Q_2) being not a sufficient Q matrix but giving the same number of KSs: the rows of matrix Q_2 are (1 1 1), (0 1 0), (0 0 1), respectively. Q_2 also maps all of the KSs to 5 IRPs but Q_2 is not a sufficient Q matrix for the three independent attributes. The sufficient Q matrix does not produce more KSs and does not guarantee to include the reachability matrix. For this reason, we introduce definitions of a necessary Q matrix and a perfect Q matrix as below, respectively.

Definition of a necessary Q matrix (Ding et al. 2011). If the reachability matrix R is a sub-matrix of a test Q matrix Q_t, Q_t is called a necessary Q matrix for representing the cognitive model of a domain of interest.

Definition of a perfect Q matrix (Ding, Wang, & Luo 2014). If a test Q matrix Q_t is a necessary Q matrix and involves minimum number of the columns, then the test Q matrix Q_t is called a perfect Q matrix.

Tatsuoka (1995, p. 341), Samejima (1995, p. 393), and Chiu et al. (2009) pointed out that if the cognitive attributes are not assumed to be organized hierarchically (i.e., an independent hierarchy) and Q_t includes identity matrix, then the set of KSs will correspond one-to-one to the set of IRPs. But the Q-matrix that Tatsuoka (1995) cited usually does not have this property. Although Leighton, Gierl and

Hunka (2004) suggested that Q_p be used as a test blueprint, sometimes the number of the columns in Q_p is too large to use. For the dichotomous situation, if the cognitive attributes are conjunctive and not compensatory, it can be proved that the reachability matrix is a perfect Q matrix for any attribute hierarchy. Next, we give Theorem 1.

Theorem 1 (Ding et al. 2010, 2011) Given non-compensatory cognitive model and 0–1 scoring, if Q_t is a necessary Q matrix with respect to attribute hierarchy, then the Q_t is one-to-one mapping from the set of KSs to the set of IRPs.

1.3 The Theoretic Construct Validity

Since some test Q matrices are not necessary Q matrices, especially for retrofitting situations, a question is: are these Q matrices efficient for cognitive diagnosis, and how to measure their efficiency?

Definition of theoretic construct validity. Suppose that the true cognitive model for the domain of interest includes N number of KSs, the potential matrix Q_p is known, a test Q matrix, Q_t, is given, and it is a submatrix of Q_p. Q_1 is obtained from Q_t using the Augment algorithm. If the number of the columns in Q_1 is N_1, the theoretic construct validity of Q_t is $(N_1 + 1)/N$.

Example 2. (Cont. Example 1). As the number of KSs of the independent attribute hierarchy with 3 attributes equals 8, and there are 4 non-zero KSs derived from Q_2 only using the Augment algorithm based on Q_2, the theoretic construct validity of Q_2 is $(4 + 1)/8 = 5/8$.

2 Polytomous Scoring Based 0–1 Matrices

It is well known that polytomous scoring is likely to yield more diagnostic information than dichotomous scoring does.

For the polytomous scoring items, if the ideal response score is equal to product of the KS and the test Q matrix, and if the test Q matrix contains the reachability matrix R, the property of bijective mapping can also be established. The above statement may be proved using linear algebra. This section devotes to seeking a perfect Q matrix for polytomous items.

2.1 Re-Partitioned Basic Attribute Hierarchies

In order to find a structure of a perfect Q matrix for polytomous scoring cognitive diagnostic test blueprint, we re-classify the attribute hierarchies to three basic types

Fig. 1 4 attributes rhomb

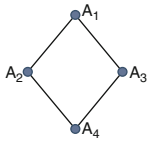


Fig. 2 5 attributes rhomb

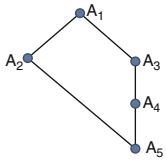


Fig. 3 convergent

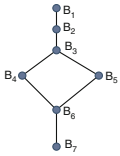
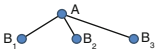


Fig. 4 Rooted tree



in terms of Graph Theory: rooted tree type that contains linear, divergent and unstructured sub-types (named by Leighton, Gierl, & Hunka 2004), independent type, and rhomb type.

Example 3. Some rhombs (Fig. 1, Fig. 2), convergent type (Fig. 3) and rooted tree (Fig. 4) for tests with polytomous scoring.

To discuss the structure of a perfect Q matrix for rhomb type, some new concepts are needed and they are introduced below.

1. If two attributes are not prerequisite to each other, these two attributes are called incomparable.
2. If there is an attribute that is the prerequisite to all other attributes in the domain of interest, the attribute is called a maximum element of the domain.
3. If there is an attribute that all other attributes in the domain of interest are its prerequisite, the attribute is called a minimum element of the domain.

Definition of a maximal set of incomparable attributes. Suppose that a set S contains all of K attributes in the domain of interest. S_1 is a non-empty subset of S. The set S_1 is called a maximal set of incomparable attributes (MSIA) in S, if S_1 satisfies all three conditions below: (a) for any attribute x in S_1, x is not a prerequisite to some (at least two) incomparable attributes, and no more than two incomparable attributes in S_1 are prerequisites to x. (b) any two attributes in S_1 are incomparable, and (c) for any attribute x in $S-S_1$ (the difference set of S and S_1), there exists an attribute in S_1, say y, that is comparable with x.

For the basic attribute hierarchical structure, an MSIA may be determined using only conditions (b) and (c) in the above definition of MSIA.

Note that there may be several different sets of MSIAs in S, but the number of the attributes in every MSIA in S must be the same.

One-to-one mapping from the set of KSs to the set of IRPs is called an optimal test blueprint. The concept of MSIA is very important to construct the optimal test blueprint for cognitive diagnosis with polytomous scoring.

Definition of a rhomb. A structure of an attribute hierarchy is called a rhomb if it satisfies following three conditions:

(a) There are a maximum element and a minimum element,
(b) There are at least 2 attributes that are incomparable,
(c) If the minimum element is deleted, the rest of hierarchical structure changes to a rooted tree type.

The convergent type named by Leighton et al. (2004) is a mixture of linear type(s) and one or more rhomb hierarchical structure(s).

2.2 Test Blueprint Design for a Test with Polytomous Items

2.2.1 For an Independent Hierarchy and a Rhomb Hierarchy

For an independent attribute hierarchy, the perfect Q matrix is a K-order nonsingular matrix. It is different from the result for 0–1 scoring situations.

Suppose a rhomb that the maximum element (say, u) and t attributes (say, v_1,v_2,…,v_t) are immediate prerequisites to the minimum element (say, w). Let $V=\{v_1,v_2,\ldots,v_t\}$ be the MSIA. The perfect Q matrix to the rhomb type is a K-by-t matrix. The structure of the perfect Q matrix is as follows: all of the elements in one column of Q matrix are 1, and the rest $t-1$ columns represent the paths from u to arbitrary $t-1$ attributes of v_1, v_2, …, v_t, respectively.

2.2.2 For Rooted Tree Hierarchy

For the rooted tree hierarchy with t leaves, the corresponding perfect Q matrix may be one of following forms: it contains all different columns that represent all of the roads from the root node (attribute) to the leaves (attributes), or it contains $(t-1)$ columns constructed from any $(t-1)$ roads from the root node to the leaves, and rest columns either constructed with all elements being 1,or constructed from the rest roads from the root to the leaves as well as any other added (Boolean add) roads from the $(t-1)$ roads.

Example 4. The rooted tree as listed in Fig. 4 and the perfect Q matrix for it.

$$Q_t^* = \begin{bmatrix} 1\ 1\ 1 \\ 1\ 0\ 0 \\ 0\ 1\ 0 \\ 0\ 0\ 1 \end{bmatrix}, \quad \text{or} \quad Q_t^{(1)} = \begin{bmatrix} 1\ 1\ 1 \\ 1\ 1\ 0 \\ 1\ 0\ 1 \\ 1\ 0\ 0 \end{bmatrix}, \quad \text{or} \quad Q_t^{(2)} = \begin{bmatrix} 1\ 1\ 1 \\ 1\ 1\ 0 \\ 1\ 0\ 0 \\ 1\ 0\ 1 \end{bmatrix}$$

$$\text{or} \quad Q_t^{(3)} = \begin{bmatrix} 1\ 1\ 1 \\ 1\ 0\ 0 \\ 1\ 1\ 0 \\ 1\ 0\ 1 \end{bmatrix}$$

2.2.3 Outline of Proofs for the Results About the Design of Perfect Matrices

The proof of the result for the independent hierarchical type is simpler. Since the rhomb hierarchical type may be considered as a rooted tree adding one minimum element, its proof may be derived from the proof of the result for the rooted tree type. The proof of the latter is rather complex. It is omitted here for brevity.

3 Polytomous *Q* Matrix Theory

3.1 Quasi-Reachability Matrix R_p

A polytomous Q-matrix is a Q-matrix whose elements are non-negative integers. In general, the results for a 0–1 Q-matrix cannot be applied directly to a polytomous Q matrix.

3.1.1 How to Obtain Polytomous Rp

Suppose that there is a $K \times K$ reachability matrix R and the highest levels for each attribute are $w_1, w_2, \ldots, w_K$, respectively. The diagonal elements of R, r_{jj}, are replaced by a vector $(1, 2, \ldots, w_i)$, each of other elements of the j column of R, say r_{ij}, is replaced by a row-vector with w_i columns, $(r_{ij}, r_{ij}, \ldots, r_{ij})$, $i, j = 1, 2, \ldots, K$. Polytomous matrix R_p is obtained with K rows, w columns,where $w = \sum_{j=1}^{K} w_j$.

Example 5. There are 3 attributes A_1, A_2 and A_3. The highest level for A_j is $4-j$, $j = 1, 2, 3$. And attribute A_1 is prerequisite to A_2 and A_3.

$$R = \begin{pmatrix} 1\,1\,1 \\ 0\,1\,0 \\ 0\,0\,1 \end{pmatrix} \rightarrow \begin{pmatrix} 1\,2\,3 & 1\,(1\,1) & 1 \\ 0\,(1\,1\,1) & 1\,2 & 0 \\ 0\,(1\,1\,1) & 0\,(1\,1\,) & 1 \end{pmatrix} = \begin{pmatrix} 1\,2\,3\,1\,1\,1 \\ 0\,0\,0\,1\,2\,0 \\ 0\,0\,0\,0\,0\,1 \end{pmatrix} = R_p$$

3.1.2 Relationship Between R_p and a 0–1 Reachability Matrix M

3.1.2.1 Expansion Algorithm

Let $e_i = (1, 0, \ldots, 0)^T$ be a column vector with w_i rows, $i = 1, 2, \ldots, K$. Partition R_p to K row-blocks and K column-blocks, and jth column-block contains w_j columns, j=1, 2, …, K. The jth row and jth column-block is replaced by w_j-order upper triangular 0–1 matrix whose elements are zeros below diagonal line of the matrix and ones elsewhere. Other elements in the jth row-block are multiplied by e_j. J=1, 2, …, K.

Using the expansion algorithm, the K-rows by w-columns polytomous matrix Q_p is changed to a w-order dichotomous matrix, denoted as M. It can be proven that M is a partial relation matrix. Therefore, M is a reachability matrix. It means that the expansion algorithm converts the polytomous matrix R_p to a w-order dichotomous reachability matrix M.

Example 6. (Cont. Example 5.)

$$R_p = \begin{bmatrix} 1\,2\,3\,1\,1\,1 \\ 0\,0\,0\,1\,2\,0 \\ 0\,0\,0\,0\,0\,1 \end{bmatrix} \rightarrow \begin{bmatrix} 1\,1\,1\,1\,1\,1 \\ 0\,1\,1\,0\,0\,0 \\ 0\,0\,1\,0\,0\,0 \\ 0\,0\,0\,1\,1\,0 \\ 0\,0\,0\,0\,1\,0 \\ 0\,0\,0\,0\,0\,1 \end{bmatrix} = \mathrm{M}$$

3.1.2.2 Compression Algorithm

The matrix M as above is partitioned to K-by-K blocks with (i, j)-block being w_i rows and w_j columns, and all the w_i rows in the (i, j)-block are added to a single row. This algorithm is called compression algorithm. It converts the w-order square matrix M into a matrix with K-rows by w-columns.

Example 7. (Continue Example 6.)

$$M = \left[\begin{array}{c|c|c} \begin{matrix}1&1&1\\0&1&1\\0&0&1\end{matrix} & \begin{matrix}1&1&1\\0&0&0\\0&0&0\end{matrix} & \begin{matrix}1\\0\\0\end{matrix} \\ \hline \begin{matrix}0&0&0\\0&0&0\\0&0&0\end{matrix} & \begin{matrix}1&1&0\\0&1&0\\0&0&1\end{matrix} & \begin{matrix}0\\0\\1\end{matrix} \end{array}\right] \Rightarrow \left[\begin{array}{c|c|c} \begin{matrix}1&2&3\\0&0&0\\0&0&0\end{matrix} & \begin{matrix}1&1&1\\1&2&0\\0&0&1\end{matrix} \end{array}\right] = Rp$$

According to the expansion and the compression algorithms, there is a one-to-one mapping between the polytomous matrix Q_p and the 0–1 reachability matrix M. Therefore, the polytomous matrix Q_p is called quasi-reachability matrix.

3.2 Polytomous Augment Algorithm

If Boolean union and join operations are changed to maximizing and minimizing operations, respectively, the Augment algorithm holds, too.

Example 8. (Cont. Example 6, 7)

$$R_p \Rightarrow \left[\begin{array}{c|c|c|c|c} 1\,2\,3\,1\,1\,1 & 2\,2\,2 & 3\,3\,3 & 1\,2\,3 & 1\,2\,3 \\ 0\,0\,0\,1\,2\,0 & 1\,2\,0 & 1\,2\,0 & 1\,1\,1 & 2\,2\,2 \\ 0\,0\,0\,0\,0\,1 & 0\,0\,1 & 0\,0\,1 & 1\,1\,1 & 1\,1\,1 \end{array}\right]$$

3.3 Polytomous Q Matrix Theory

The polytomous augment algorithm has addressed the problem about the universal set of KSs. Other problems of polytomous *Q* matrix theory are settled as follows.

3.3.1 The Role of R_p in Construction of Test Blueprint

Given a knowledge state (also called the attribute master pattern) α, $\alpha = \{(\alpha_1, \cdots, \alpha_k)\}$, and an attribute vector of item j, $Q_j = \left(q_{1j} \cdots q_{kj}\right)^{\mathrm{T}}$,an ideal response of α to item j can be designed as (Sun et al. 2013)

$$S_j(\alpha) = \sum\nolimits_{k=1}^{K} q_{kj} I_{(\alpha_k \geq q_{kj})} \tag{1}$$

Theorem 2 If R_p is a submatrix of $Q_t^{(p)}$, α, β are knowledge states and $\alpha \neq \beta$, then the IRP satisfies $\alpha \circ Q_t^{(p)} \neq \beta \circ Q_t^{(p)}$.

The $Q_t^{(p)}$ that includes the quasi-reachability matrix R_p is called polytomous necessary Q matrix.

3.3.2 How to Obtain the Attribute Hierarchy From R_p

The main objective is to reduce the columns of R_p by using the deleting method: if jth column of R_p can be represented by other column(s) of R_p, then the jth column is deleted from R_p. The resulted matrix is a 0–1 reachability matrix. Using the Clean algorithm (Ding & Luo 2005), its attribute hierarchy can be derived.

4 Discussion

The dichotomous and polytomous Q matrix theory has been applied to construct some cognitive diagnostic models (Luo, Li, Yu, Gao, & Peng 2015; Sun et al. 2013; Sun, Zhang, Xin, & Bao 2011; Tu, Cai, Dai, & Ding 2012), and is also applied to discuss the attribute-level and pattern-level classification consistency and accuracy indices (Wang, Song, Chen, Meng, & Ding 2015).

The problem of test blueprint design for cognitive diagnosis and especially of its optimal design, is related to multiple factors, such as the cognitive model (the attribute hierarchies, the relation among the attributes being compensatory or not), the scoring rubric (0–1 or polytomous), the element values of the Q matrices (dichotomous or polytomous), and even the cognitive diagnostic model being chosen, because some cognitive diagnostic models are robust to the cognitive model and some are not.

When R and Q are 0–1 matrices but the scoring rubric at an item is polytomous using Eq. (1), the results of this note are true; furthermore, fewer item(s) can construct a one-to-one mapping from the set of KSs to the set of IRPs (Ding, Luo, & Wang 2014; Ding, Wang, et al. 2014). These results may be useful for classroom assessment where shorter test is required. The results obtained in this note are based on the hypothesis that the Q matrix is correct. In reality, some elements in a Q matrix may be wrong, and some researchers have paid attention to correct the Q matrix.

For compensatory cognitive model, the optimal design of cognitive diagnostic test blueprint remains a big challenge and is worthy of further investigations.

Acknowledgments This research was supported by the National Natural Science Foundation of China (Grant No. 31500909, 31360237, 31160203, 31300862, 31300876, and 61262080), the National Social Science Foundation of China (Grant No.12BYY055, 13BYY087), the Humanities and Social Sciences Research from China Ministry of Education (Grant No. 13YJC880060 and 12YJA740057). Thank Professor Douglas for reviewing the paper in patience. Thank Dr. Guan Xiaosheng for revising the paper. Those help me to increase the readability of the paper.

References

Chen, J. S., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement, 37*(6), 419–437.

Chiu, C., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*(4), 633–665.

Ding, S. L., & Luo, F. (2005). Algorithm: From poset to Hasse diagram. *Journal of Jiangxi Normal University (Natural Science Edition), 29*(2), 150–152.

Ding, S. L., & Luo, F. (2013). An efficient algorithm of driving Hasse diagram from the reachability matrix of a partial relation—Together with its application to cognitive diagnosis. *Journal of Jiangxi Normal University (Natural Science Edition), 37*(5), 441–444.

Ding, S. L., Luo, F., Cai, Y., Lin, H. J., & Wang, X. B. (2008). Complement to Tatsuoka's Q matrix theory. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 417–423). Tokyo, Japan: Universal Academy Press, Inc.

Ding, S. L., Luo, F., Wang, W. Y., & Xiong, J. H. (2015). The properties of 0–1 and polytomous reachability matrices and their applications. *Journal of Jiangxi Normal University (Natural Science Edition), 39*(1), 64–68.

Ding, S. L., Luo, F., & Wang, W. Y. (2012). Extension to Tatsuoka's Q matrix theory. *Psychological Exploration, 32*(5), 410–422.

Ding, S. L., Luo, F., & Wang, W. Y. (2014). Design of polytomous cognitively diagnostic test blueprint—For the independent and the rhombus hierarchies. *Journal of Jiangxi Normal University (Natural Science Edition), 38*(3), 265–269.

Ding, S. L., Wang, W. Y., & Luo, F. (2014). Design of polytomous cognitively diagnostic test blueprint—For the rooted tree type. *Journal of Jiangxi Normal University (Natural Science Edition), 38*(2), 111–118.

Ding, S. L., Wang, W. Y., & Yang, S. Q. (2011). The design of cognitive diagnostic test blueprints. *Journal of Psychological Science, 34*(2), 258–265.

Ding, S. L., Yang, S. Q., & Wang, W. Y. (2010). The importance of reachability matrix in constructing cognitively diagnostic testing. *Journal of Jiangxi Normal University (Natural Science Edition), 34*(5), 490–495.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: a variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*(3), 205–237.

Luo, Z. S., Li, Y. J., Yu, X. F., Gao, C. L., & Peng, Y. F. (2015). A simple cognitive diagnosis method based on Q-matrix theory. *Acta Psychological Sinica, 47*(2), 264–272.

Rosen, K. H. (2003). *Discrete mathematics and its applications* (5th ed.). Beijing, China: China Machine Press.

Samejima, F. (1995). A cognitive diagnosis method using latent trait models: Competency space approach and its relationship with DiBello and Stout's unified cognitive-psychometric

diagnosis model. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessments* (pp. 391–410). Hillsdale: Erlbaum.

Sun, J. N., Xin, T., Zhang, S. M., & de la Torre, J. (2013). A polytomous extension of the generalized distance discriminating method. *Applied Psychological Measurement, 37*(7), 503–521.

Sun, J. N., Zhang, S. M., Xin, T., & Bao, Y. (2011). A cognitive diagnosis method based on Q matrix and generalized distance. *Acta Psychological Sinica, 43*(9), 1095–1102.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern classification approach. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessments* (pp. 327–359). Hillsdale: Erlbaum.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York, United States: Routledge, Taylor & Francis Group, LLC.

Tu, D. B., Cai, Y., Dai, H. Q., & Ding, S. L. (2012). A new multiple-strategies cognitive dignosis model: The MSCD method. *Acta Psychological Sinica, 44*(11), 1547–1553.

Wang, W. Y., Song, L. H., Chen, P., Meng, Y. R., & Ding, S. L. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement, 52*(4), 457–476.

Yang, S. Q., Cai, S. Z., Ding, S. L., Lin, H. J., & Ding, Q. L. (2008). Augment algorithm for reduced Q matrix. *Journal of Lanzhou University (Natural Sciences), 44*(3), 87–91. +96.

Yang, S. Q., & Ding, S. L. (2011). Theory and method for predicating of valid objects. *Journal of Jiangxi Normal University (Natural Science Edition), 35*(1), 1–4.

Zhang, S. M. (2012). *Researches on new models of polytomous cognitive diagnosis*. Unpublished Doctoral Thesis, Beijing Normal University.

Multidimensional Joint Graphical Display of Symmetric Analysis: Back to the Fundamentals

Shizuhiko Nishisato

Abstract The basic premise of dual scaling/correspondence analysis lies in the simultaneous or symmetric analysis of rows and columns of the data matrix, a task that resembles the analysis of principal component analysis of both the person-to-person correlation matrix and the item-by-item correlation matrix together. Our main quest: whether or not we can represent both analyses in the same Euclidean space. The traditional graphical methods are very problematic: symmetric display or French plot suffers from the discrepancy between the row space and the column space; non-symmetric display involves the projection of data onto standardized space, which does not contain coordinate information in the data; a variety of biplots, of which criticisms we rarely see, involve operations that do not typically maintain row and column measurements on the equal metrics, or if they do they are not the coordinates of the data. Thus, none of these provides a precise description of complex information in data, hence failing in the basic objective of symmetric data analysis. This paper will identify logical problems of the current practice and offers a justifiable alternative to joint graphical display. "Graphing is believing" may in reality remain to be a wishful thinking.

Keywords Duality • Joint space for rows and columns • Doubling multidimensional space

1 Introduction

This paper deals with graphical display of quantification theory, where the main interest lies in the joint analysis of rows and columns of the data matrix. This aspect is reflected by the word 'dual' of Canadian dual scaling (Nishisato 1980) used to treat rows and columns of a data matrix on the equal footing, that is, symmetric analysis of the data matrix. The technique is referred to by many other names such as British simultaneous linear regressions (Hirschfeld 1935), the

S. Nishisato (✉)
University of Toronto, Toronto, ON, Canada
e-mail: shizuhiko.nishisato@utoronto.ca

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer
Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_22

American method of reciprocal averages (Horst 1935), Hayashi's Japanese theory of quantification (1950), American principal component analysis of categorical data (Torgerson 1958), American optimal scaling (Bock 1960), French 'analyse des correspondances' (Escofier-Cordier 1969), and Dutch homogeneity analysis (De Leeuw 1973). See many other names in Nishisato (2007).

In the traditional multivariate analysis, we often use the least-squares procedure, which means projection of, for example, data onto the model space, meaning a one-directional analysis as opposed to the two-way symmetric analysis of equal norms. Graphical display of quantification results must be such that the norm of the row variables should be equal to the norm of the column variables. This is a difficult task for joint graphical display of quantification theory, and in the past a number of methods have been proposed, none of which, however, is satisfactory. The current paper starts with some basic premises of quantification, and then discusses how the perennial problem of joint graphical display should be dealt with. We start with some relevant basic points.

2 Fundamental One: Orthogonal Coordinates for *n* Variables

When we wish to show a graph of two sets of scores (e.g., Mathematics test and language test), it is a widely used practice to introduce the horizontal axis for the mathematics test and the vertical axis for the language test as if the two variates were orthogonal to each other. This is definitely wrong, but this practice has been used widely for many years. When we have a number of variables, say n, where $n > 1$, the first task for graphical display is to introduce an orthogonal coordinate system to accommodate these variables. There are an infinite number of such systems, and the most widely used choice, out of them, is to adopt principal coordinates, through principal component analysis: Given the subject-by-test data matrix, $\mathbf{F}$, we calculate the test-by-test correlation matrix R, which is then subjected to the eigenvalue decomposition, that is, $\mathbf{R} = \mathbf{X}'\boldsymbol{\Delta}\mathbf{X}$, where $\mathbf{X}$ is the subject-by-test matrix of coordinates and $\boldsymbol{\Delta}$ is the diagonal matrix of eigenvalues. The number of non-zero elements of $\boldsymbol{\Delta}$ is the required dimensionality of the space for multidimensional coordinates of n variables.

3 Fundamental Two: Coordinates of Framework and Variables

In this principal axis decomposition of data matrix $\mathbf{F}$, $\mathbf{X}$ is referred to as the matrix of *standard coordinates* and $\boldsymbol{\Delta}^{1/2}\mathbf{X}$ is called the matrix of *principal coordinates*. It is crucial for graphical display to distinguish between these two coordinates. Nishisato (1996) explained the important difference between them using a simple example

as follows: Consider principal component analysis of standardized variables, and suppose that the data are two-dimensional, then plotting principal coordinates of variables results in a perfect circle with the diameter 1, where all data points lie; suppose that the data are perfectly three-dimensional, then plotting the principal coordinates of the data reveals that all data points lie at a distance of 1 from the origin on the three-dimensional sphere, or on the perfect ball. If we plot standard coordinates, instead of principal coordinates, however, the two-dimensional data will show, not a perfect circle, but typically an elongated circle. If the first eigenvalue is comparatively larger than the second one, the graph will be elongated toward the second dimension. In other words, standard coordinates do not describe the structure of the data, but a function of the distribution of data under the condition that the sum of squares on each dimension is constant, thus the name standard (*i.e.,*the fewer the responses the larger the standard coordinates). The conclusion here is that the coordinates of variables in multidimensional space are given by principal coordinates.

4 Fundamental Three: Dual Relations

Quantification theory can be depicted as singular value decomposition of data matrix **F**, that is, $\mathbf{Y\Lambda X}'$, where **Y** and **X** are standard coordinates of rows and columns, respectively, and $\mathbf{\Lambda}$ is the diagonal matrix of singular values. Because of the symmetry of this analysis, Nishisato (1980) called it dual scaling, based on the dual relations:

$$\rho_k y_{ik} = \frac{\sum_{j=1}^{m} f_{ij} x_{jk}}{f_i} \quad \text{and} \quad \rho_k x_{jk} = \frac{\sum_{i=1}^{n} f_{ij} y_{jk}}{f_j}$$

where ρ_k is the k-th singular value, f_{ij} is the element of the i-th row and the j-th column, $f_{i.}$ and $f_{.j}$ are respectively the sums of the i-th row and that of the j-th column of data matrix **F**. In other words, for each component k, the mean of rows i of **F**, weighted by column weights x_j is equal to the weight for row i times the singular value, and the mean of column j, weighted by row weights y_i is equal to the weight for column j times the singular value. This mutual reciprocal averaging relation holds for each component. Although ρ_k is the singular value of data matrix **F**, it is also (1) Hirschfeld's simultaneous regression coefficient (1935), (2) Guttman's maximal row-column correlation (1941) and (3) Nishisato's (1980) projection operator from row space to column space or vice versa.

5 Fundamental Four: Discrepancy Between Row Space and Column Space

For a particular component, the dual relation shows that the mean of the row i, weighted by column weights x_j, is equal to the weight for row i times the singular value. In other words, the singular value is the projection operator of the row space onto the column space, or vice versa. Thus, it is possible to calculate the angle of discrepancy, θ_k, between the row space and the column space for component k by the following formula (Nishisato & Clavel 2008):

$$\theta_k = \cos^{-1}\rho_k$$

From this we know that only when the singular value is one the variables associated with rows and columns of the data matrix span the same space. In this regards, we should remember the famous warning by Lebart, Morineau and Tabard (1977) that one cannot calculate the exact distance between a row variable and a column variable from the symmetric scaling.

6 Lessons From Analysis of Contingency Table and Response-Pattern Table

Using an example from Nishisato (1980), some important aspects of joint graphical display can be illustrated to clarify the current controversies of joint graphical display.

Consider the following 2×3 contingency table, $\mathbf{C}$, obtained by asking two multiple-choice questions:

Q1: Do you smoke? (yes, no)
Q2: Do you prefer coffee to tea? (yes, not always, no)

Suppose we obtained the following data indicated by $\mathbf{C}$, which is the 'options of Q.1-by options of Q.2,' that is, a 2×3 table of joint frequencies. Nishisato (1980) has shown that the same data can be represented also as the traditional response-pattern table $\mathbf{F}_a$, which is the 'subjects-by-options of two items,' that is, 14×5 incidence table. He has also shown that this large table can be transformed into a condensed response-pattern table $\mathbf{F}_b$ by creating a table of distinct patterns with frequencies. In our example, the data in the three data formats are as follows:

$$\mathbf{C} = \begin{bmatrix} 3 & 2 & 1 \\ 1 & 2 & 4 \end{bmatrix}; \quad \mathbf{F}_a = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix}; \quad \mathbf{F}_b \begin{bmatrix} 3 & 0 & 3 & 0 & 0 \\ 2 & 0 & 0 & 2 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 2 & 0 & 2 & 0 \\ 0 & 4 & 0 & 0 & 4 \end{bmatrix}$$

As Nishisato (1980) has shown, the two response-pattern formats yield identical quantification results. Therefore, for brevity we will use **F**.

Suppose that two items have n and m options, respectively, and there are N respondents. Then, assuming that N is much larger than the sum of the response categories, the total number of components from the $n \times m$ contingency table, $K(\mathbf{C})$, is equal to the smaller of n and m minus 1, that is,

$$K(\mathbf{C}) = \min(n, m) - 1.$$

In the current example, min(2,3)−1 = 2−1 = 1. Assuming that N is much larger than $n + m$, the total number of components from the response-pattern table, $K(\mathbf{F})$, is equal to the total number of categories of two items minus 2, that is,

$$K(\mathbf{F}) = n + m - 2.$$

In the current example, $K(\mathbf{F}) = 2 + 3 - 2 = 3$.

According to the Young-Householder theorem (Young & Householder 1938), the variates within columns (or, rows) of the data matrix can be mapped in the same Euclidean space. Thus, the coordinates of those five columns of **F** can be mapped in the same Euclidean space. In contrast, we have already shown that the two rows and the three columns of **C** do not belong to the same space. From this comparison, we can draw the conclusion that the five options of the two items require three-dimensional space to be plotted together. Our numerical example (Table 1) yields the following coordinates on respective dimensions. Notice that the standard coordinates associated with **C** are exactly the same as the standard coordinates of the corresponding first component of **F**:

Several years after Nishisato's book was published, Carroll, Green and Schaffer (1986) wrote a paper on the method called the CGS scaling, in which they maintained that the space discrepancy between row and column space of the

Table 1 Standard coordinates associated with the two formats of data

	The results of **C**	The results of $\mathbf{F}_b$		
Component	1	1	2	3
Smoking 'yes'	1.08	1.08	0.00	−1.08
Smoking 'no'	−0.93	−0.93	0.00	0.93
Coffee 'yes'	1.26	1.26	−1.14	1.26
'not always'	0.17	0.17	2.11	0.16
'no'	−1.14	−1.14	−0.77	1.14

contingency table could be solved by representing the rows and the columns of the contingency table into the same columns of the response-pattern table—this is exactly what was shown above. However, the CGS scaling was severely criticized by Greenacre (1989) as false, and his criticism resulted in the downfall of the CGS scaling. What these investigators completely missed was the point that the weights for the rows and those for the columns of the contingency table require more dimensions if they are represented in the same rows of the response-pattern table. In the above example, one needs three dimensions. In the above example, the singular value of the component associated with the contingency table is 0.4590, thus the discrepancy angle between the row axis and the column axis is 62.68°, leading to the conclusion that we need more than one dimension for the data. The idea of the CGS scaling should have been presented under the condition that the space dimensionality must be at least doubled from that of the contingency table.

7 Dimensionality of Total Space

In the above comparison of the contingency format and the response-pattern format, we concluded that those response options of the two items can be mapped in the same space, provided that the dimensionality of the space is expanded. There are two distinct views on how many dimension are needed. The first one is Nishisato's view of doubled multidimensional space (2012). His idea of 'doubling' comes from the consideration that for each component we must introduce two axes with the angle of $\cos^{-1}\rho_k$. His view looks reasonable, but we need another view on this: Based on the comparison between quantification of the contingency table and that of the corresponding response-pattern table, we need to double the dimensionality or more than double the dimensionality. This view stems from the following fact:

$K(\mathbf{F}) = 2 \times K(\mathbf{C})$, when $n = m$ and
$K(\mathbf{F}) > 2K(\mathbf{C})$, when $n \neq m$

In other words, only when the number of options of Item 1 is equal to that of Item 2, we need to double the dimensionality. Otherwise, as was the case of the above numerical example, we need more than double the dimensionality of the joint space.

8 From Joint Graphical Display to Cluster Analysis of Total Space

Nishisato (1997) wrote a paper on "Graphing is believing" in support of graphical display. With the current revelation, however, it seems generally impossible to summarize data in multidimensional space, for we are limited to grasp or understand only two- or three-dimensional graphs and the total space for the joint graphical display with principal coordinates is almost always greater than two or three. At this juncture, Nishisato and Clavel (2010) proposed total information analysis or comprehensive dual scaling: Extract all components from the data, calculate the within-row distance matrix, the between row-column distance matrix and the within-column distance matrix; subject this super-distance matrix to cluster analysis, to identify clusters in the total space as defined here. In this way, we do not have to concentrate only on major configurations, but can also look at other rare combinations of variables. (see Nishisato (2014) for a numerical example.) Total information analysis has not widely been applied to data analysis yet, but is definitely a logical and reasonable alternative to the traditional analysis via multidimensional joint graphical display.

9 Concluding Remarks

Historically, French correspondence analysis placed a major emphasis on joint graphical display. The current paper has identified a number of logical problems associated with joint graphical display, be it symmetric French plot, or non-symmetric plot, or biplot. A number of those logical problems prompted Nishisato and Clavel (2010) to propose total information analysis (TIA), which as explained in the current paper is free from any logical problems. It is hoped that through many applications of TIA to data we will learn further how practical and useful TIA is as an alternative to the traditional multidimensional joint graphical approach to data analysis.

References

Bock, R. D. (1960). Methods and applications of optimal scaling. *The University of North Carolina Psychometric Laboratory Research Memorandum*, No. 25.

Carroll, J. D., Green, P. E., & Schaffer, C. M. (1986). Interpoint distance comparison in correspondence analysis. *Journal of Marketing Research, 23*, 271–280.

De Leeuw, J. (1973). *Canonical analysis of categorical data.* Unpublished doctoral dissertation, Leiden University, Leiden, The Netherlands.

Escofier-Cordier, B. (1969). L'analyse factorielle des correspondances. *Bureau Universitaire de Recherche Operationnele. Cahiers, SérieRecherche* (Université de Paris), *13,* 25–29.

Greenacre, M. J. (1989). The Carroll-Green-Schaffer scaling in correspondence analysis: A theoretical and empirical appraisal. *Journal of Marketing Research, 26*, 358–365.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In P. Horst et al. (Eds.), *The Prediction of Personal Adjustment* (pp. 319–348). New York, NY: Social Research Council.

Hayashi, C. (1950). On the quantification of qualitative data from mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics, 2*, 35–47.

Hirschfeld, H.O. (1935). A connection between correlation and contingency. *Cambridge Philosophical Society Proceedings, 31,* 520–524.

Horst, P. (1935). Measuring complex attitudes. *Journal of Social Psychology, 6*, 369–374.

Lebart, L., Morineau, A., & Tabard, N. (1977). *Tecnniques de la description statistique: Méthodes et l'ogiciels pour l'analyse des grands tableaux*. Paris, France: Dunod.

Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto, ON, Canada: University of Toronto Press.

Nishisato, S. (1996). Gleaning in the field of dual scaling. *Psychometrika, 1996*(61), 559–599.

Nishisato, S. (1997). Graphing is believing: Interpretable graphs for dual scaling. In J. Blasius & M. J. Greenacre (Eds.), *Visualization of categorical data* (pp. 185–196). London, England: Academic.

Nishisato, S. (2007). *Multidimensional nonlinear descriptive analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Nishisato, S. (2012). Quantification theory: Reminiscence and a step forward. In W. Gaul, A. Geyer-Schults, I. Schmidt-Thieme, & J. Kunze (Eds.), *Challenges and the interface of data analysis, computer science and optimization* (pp. 109–119). New York, NY: Springer.

Nishisato, S. (2014). Structural representation of categorical data and cluster analysis through filters. In W. Gaul, A. Guyer-Schultz, A. Baba, & A. Okada (Eds.), *German-Japanese interchange of data analysis results* (pp. 81–90). New York, NY: Springer.

Nishisato, S., & Clavel, J. G. (2008). A note on between-set distances in dual scaling and correspondence analysis. *Behaviormetrika, 30*, 87–98.

Nishisato, S., & Clavel, J. G. (2010). Total information analysis: Comprehensive dual scaling. *Behaviormetrika, 37*, 15–32.

Torgerson, W. S. (1958). *Theory and Methods of Scaling*. New York, NY: Wiley.

Young, G., & Householder, A. A. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika, 3*, 19–22.

Classification of Writing Patterns Using Keystroke Logs

Mo Zhang, Jiangang Hao, Chen Li, and Paul Deane

Abstract Keystroke logs are a valuable tool for writing research. Using large samples of student responses to two prompts targeting different writing purposes, we analyzed the longest 25 inter-word intervals in each keystroke log. The logs were extracted using the ETS keystroke logging engine. We found two distinct patterns of student writing processes associated with stronger and weaker writers, and an overall moderate association between the inter-word interval information and the quality of final product. The results suggest promise for the use of keystroke log analysis as a tool for describing patterns or styles of student writing processes.

Keywords Keystroke logs • Writing processes • Writing pattern • Inter-word interval

1 Introduction

Keystroke logs (KL) are a valuable tool for writing research (Leijten & van Waes 2013). A keystroke logging program essentially records the mechanical processes of one's writing and the temporal information associated with an action (e.g., insert, delete). Information recorded and extracted from KLs can provide rich evidence on one's writing practice, writing proficiency, linguistic skills, as well as underlying cognitive processes (e.g., Leijten, Macken, Hoste, van Horenbeeck & van Waes 2012). There is an extensive research literature on writing using KLs. For example, van Waes, Leijten and van Weijen (2009) showed that KLs can be used for conducting empirical research on comparing writers with different ability levels, on studying cognitive processes during writing, and on examining learning styles. Many other published studies have used KLs for studying writing strategies (Xu & Ding 2014), genre effects (Beauvais, Olive & Passerault 2011), keyboarding skills (Grabowski 2008), and writing skills for native speakers vs. non-native speakers (Miller 2000; Roca de Larios, Manchon, Murphy & Marin 2008). Lastly, studies have used KLs for writing research in various contexts, such

M. Zhang (✉) • J. Hao • C. Li • P. Deane
Educational Testing Service, Princeton, NJ 08541, USA
e-mail: MZhang@ETS.org; JHao@ETS.org; CLi@ETS.org; PDeane@ETS.org

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer
Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_23

as spontaneous communication (Chukharev-Hudilainen 2014), professional writing (Leijten, van Waes, Schriver & Hayes 2014), educational assessment (Deane 2014), language translation (Dragsted & Carl 2013), and cognitive modeling (Almond, Deane, Quinlan & Wagner 2012). This literature indicates that stronger writers tend to write more fluently and persistently, spending more time on task, writing in longer "bursts" of uninterrupted text production, and pausing primarily at locations such as sentence and clause boundaries that indicate an emphasis on sentence and discourse-level planning. These results have been confirmed in an assessment context in our own studies (Deane 2014; Deane & Zhang 2015; Zhang & Deane 2015).

From an assessment perspective, evaluating writing processes can be valuable for at least two reasons. One is to give teachers and students (diagnostic) feedback to improve writing practice. As a simple example, a limited time on task might suggest a lack of persistence. Similarly, an absence of editing behavior, in combination with disjoint text (e.g., containing substantial amount of grammatical and mechanical errors interfering the flow and meaning of the text), might prompt attention to teaching revision strategies (Zhang & Deane 2015). A second potential value is to characterize differences among subpopulations beyond the quality of the final product, such as mechanical characteristics of text production (e.g., typing speed, extent of editing) and writing patterns and styles (e.g., lack of planning). It is this latter idea that is the focus of this study.

We used a KL engine developed at Educational Testing Service, which can produce a large number of features such as the length of a within-word pause (Deane 2014). In this study, we focused on the inter-word interval (IWI) feature, or pauses between two adjacent words, extracted from the KL engine in order to address two research questions: (1) Can we distinguish writing patterns using this information? (2) How does this information relate to the human ratings of text-production skills (e.g., grammar, organization)?

Previous research suggests that IWIs tend to be associated with such cognitive activities for word and sentence planning and deliberation, especially when lengthy and at sentence, clause, or discourse boundaries (Baaijen, Galbraith & de Glopper 2012; Banerjee, Feng, Kang & Choi 2014; Chenoweth & Hayes 2001; Chukharev-Hudilainen 2014; Gould 1980; Xu & Ding 2014), though IWIs can also indicate difficulties in keyboarding and lexical access such as in spelling. Shorter IWIs, in contrast, are more likely to reflect basic keyboarding fluency (e.g., Alves, Castro & de Sousa 2007).

Figure 1 provides a dendrogram visualization of the IWI duration using one essay as an example. The original essay reads as below, with bolded emphasis added by the authors.

> I think the community should build a swimming pool because i believe you can **I maintain** a healthier lifestyle and have fun while doing **it It** has more pottential to benefit more people than than a foreign **exchange. The** whole community can benefit the pool but only a select few would benefit from the exchange program.

At the bottom of the figure are the words produced. The height of the horizontal lines that connect two words (i.e., the ones with the shortest distance on the time

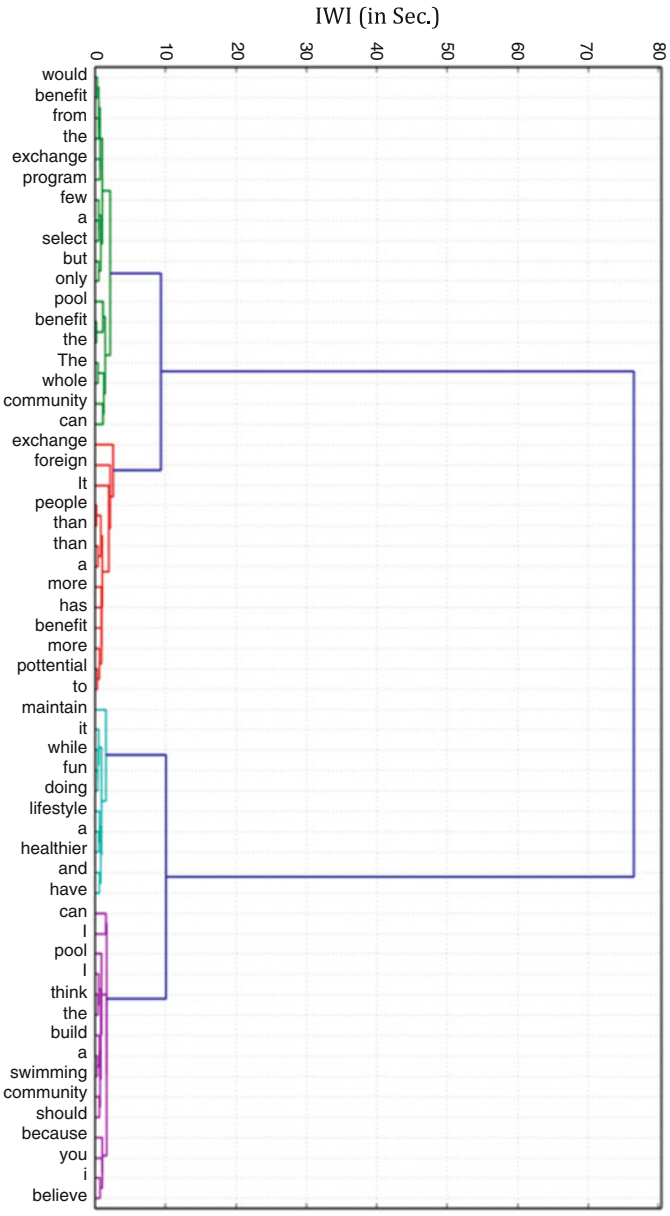


Fig. 1 Visualizing the IWI duration

dimension from two clusters) represents the IWI duration in seconds. We note that, in this example, the three longest between-word pauses all happen at the phrasal and sentence boundaries, which naturally separate the text production into four bursts. Specifically, the pause between "exchange" and "The" is the longest with more than 70 s, with "exchange" on the end of the previous sentence and "The" being the first word of the following sentence. It is reasonable to assume that the student spent a fair amount of time planning the second sentence or reading and evaluating the source materials.

The pauses between the words "I" and "maintain" (where "I" is likely to be a typo), and the words "it" and "It," both occur at reasonable junctures between long chunks, each pause lasting for about 10 s. A sensible explanation for these relatively long pauses is that the student was deliberating on the choice of words and sentence structure. All the other IWIs in this example were less than 2 or 3 s.

2 An Approach to Comparing Keystroke Logs

Even though the IWI can provide temporal evidence about the essay composition process, we need to quantify and summarize this information in a way that can be used to make meaningful distinctions among groups of students. A practical challenge is how to align the IWIs from different keystroke logs so that we can compare overall writing patterns despite significant differences in the time individual students spend on writing and the number of words they ultimately produce.

In this study, we explore one novel method of comparison, which attempts to preserve information about the longer pauses (likely to reflect strategic processing) and to interpret them in terms of their temporal position in the student's writing process. A software implementation of this method can be found in Hao, Smith, Mislevy, von Davier and Bauer (2016). Our attempt in this study focused on the longer between-word pauses (IWIs). First, for each log, we rank the IWIs from the longest to the shortest. Then, we choose the length (or duration) of IWI and its normalized median time point along the time axis as two indicative variables to represent each keystroke log. By placing these indicative variables one after another based on the rank ordering of the IWI duration, we form a vector of IWI for each keystroke log. Subsequently, for all vectors (from different logs) to have the same length, we need to truncate each vector by introducing a cut-off in the rank ordering. In our implementation, we choose 25 longest IWIs as the cut-off rank, for which rationale is to be explained shortly.

Through this aligning procedure, to this point, each keystroke log is represented by an IWI vector with 50 elements that capture the duration and temporal location (i.e., median time-points) of the 25 longest IWIs We can then compare the similarity or difference among logs using the IWI vectors. Finally, it is worth mentioning that, to align keystroke logs of different length in time, we normalized the median time-points of the IWIs by the total writing time to the range between zero and one.

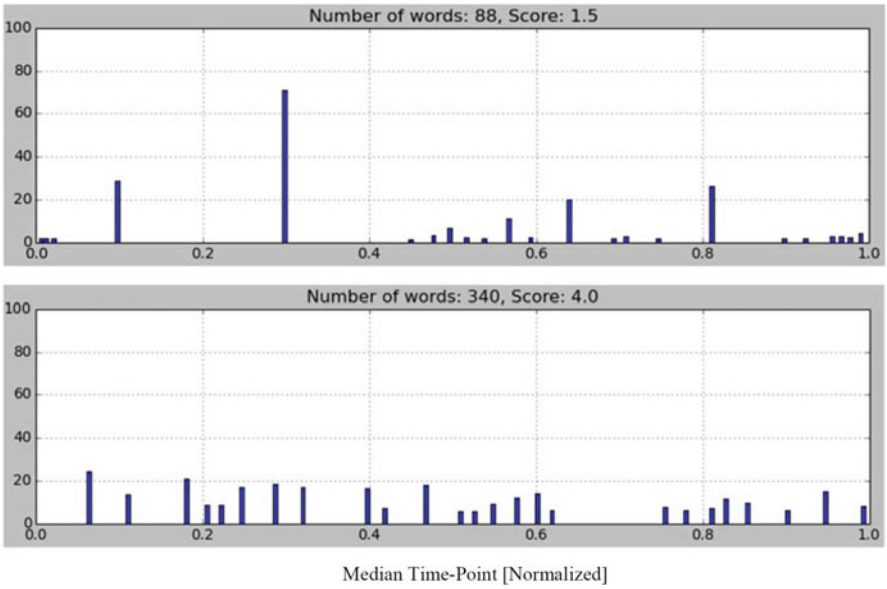


Fig. 2 Visualizing the top 25 IWI durations and median time-points for two essays written by different students

The decision to restrict attention to the longest 25 IWIs is motivated by the overall distribution of IWIs, which follows a heavily skewed distribution in which almost all IWIs are very short (less than half of a second). The psycholinguistic literature referenced above suggests rather different interpretations for the main part of the distribution (which reflects fluent text processing), and the tail consisting of very long pauses (which are more likely to reflect strategic processes such as discourse-level and sentence-level planning). Also, the cut-off rank of 25 is the largest number that essays can reach in our dataset before cases need to be excluded.

Figure 2 gives two examples of different students' writing processes, each represented by the longest 25 IWIs and their median time-points.

The vertical axis indicates the length of IWIs in seconds, and horizontal axis gives the corresponding median time-point of an IWI. The median time-point is computed as the mid-point of the absolute time between two adjacent words. For example, if the last character of the first word is typed in minute 1 second 30 and the first character in the following word is typed in minute 1 second 40, the median time-point of the IWI is minute 1 second 35. Also provided in the graphs are the total number of words in the final product and the human score on a rubric of basic writing skills for each essay.

As can be seen, the two individual KL examples in Fig. 2 are drastically different. The top panel shows that this student has made few relatively long inter-word pauses, with four instances standing out as particularly extensive (i.e., greater than 20 s). The lower panel shows a very different pattern of IWI duration and median

time-point derived from a different essay. Compared to the example above, the longest 25 IWIs in the lower panel are somewhat more evenly distributed throughout the composition and the length of the durations seems to vary less across the 25 IWIs. By normalizing the median time-point by the total writing time, we removed the confounding factor of time on task (which also relates to the essay length). Figure 2 suggests that the IWI information provides a different type of evidence about writing proficiency from the essay length and human scores on essay quality.

3 Method

3.1 Instrument

We used two summative ELA (English Language Arts) writing assessments developed by the Cognitively Based Assessment *of*, *for*, and *as* Learning (CBAL™) research initiative at Educational Testing Service (ETS®) for this investigation (Bennett 2010, Bennett, Deane, and van Rijn 2016). The CBAL scenario-based assessments are designed to have a progression from three sections of lead-in tasks to a culminating essay task. Deane et al. (2015) describe the theoretical foundation for the design of this test structure in detail.

In this study, we used one assessment from each of two writing purposes: Policy Recommendation and Argumentation. The targeted level was grade 7 for policy recommendation and grade 8 for argumentation. For our investigation, we focused on the essay task only. Specifically, the policy recommendation essay asks students to evaluate two proposals (i.e., on how to spend a generous monetary donation to the school) and write an essay recommending one policy over the other. The argumentation essay asks students to take a position using reasons and evidence as to whether students should be rewarded for getting good grades. Students are provided with a planning tool and can access three source materials at any time (that were used in the lead-in tasks). Students are encouraged to utilize the examples and viewpoints given in the source materials (with appropriate quoting or paraphrasing).

The essays were graded on two scoring rubrics: writing fundamentals and higher-level skills targeted at the specific writing purpose. The human score scale is integer 0–5 for both rubrics, where 0 indicates some unusualness in the essay response (e.g., empty, off topic). Responses were graded by two independent teacher raters who were trained by ETS assessment specialists, with a possible third rater for adjudication. We evaluated the inter-rater agreements using quadratically weighted kappa (QWK) and percentage agreement. Consider the two human scores (x and y) form a 2-by-2 matrix, and the QWK is computed as $k = 1 - \sum_{xy} w_{xy} O_{xy} / \sum_{xy} w_{xy} E_{xy}$, where w_{xy} is the weighting matrix, O_{xy} is the observed pair of rating E_{xy} is expected pair of rating. Further, the weight follows a quadratic function of the differences between the two scores; that is, the weight w_{xy} equals to 0, 1, 4, 9, and 16 for score differences of 0, 1, 2, 3, and 4, respectively (Fleiss & Cohen 1973). QWK can range

from 0 to 1 with 0 indicating no association between the two ratings and 1 indicating perfect association between the two ratings. The exact percent agreement is the sum of the percentages of the pair of ratings where the ratings are exact the same (i.e., on the diagonal in the cross-tab table). The one-point adjacent percent agree is the sum of percentages of the pair of ratings where the two human scores are within one score point. Both exact and one-point adjacent percent agreement can range from 0 to 100, with 0 indicating no agreement between the two ratings and 100 indicating perfect association between the two ratings.

The first and second rater agreements were a QWK of 0.63, exact percent agreement of 51, and one-point adjacent percent agreement of 97 for the policy recommendation essay. The comparable values were 0.67, 52, and 97 for the argumentation essay. For our analyses, we used adjudicated human scores as the criterion variable (to address Research Question 2). That means when the first two human scores were within one point apart, we used the average of the two human scores. When the first two human scores were discrepant by more than one point, we used the average of the two closest scores. If all three scores were two points apart, we used the middle score.

3.2 Data Set

The data set was collected as part of a larger study conducted in multiple US states in 2013. Included in this study were 831 essay responses from the policy recommendation assessment, and 902 essay responses from the argumentation assessment. For the policy recommendation assessment, 36 % of the participants were male, 35 % female, and the remaining 29 % unreported. As for ethnicity, nearly half of the participants (49 %) were Caucasian; 16 % were Hispanic; 6 % were Asian, African American, Middle Eastern, mixed, or others; and 29 % did not indicate. Finally, most of the participants in this assessment were seventh graders (52 %), 9 % were sixth graders, 10 % were eighth graders, and the remaining 29 % provided no information. For the argumentation assessment, 42 % of the participants were female, 38 % male, and 20 % did not report on their gender. For ethnicity, 60 % were Caucasian, 13 % Hispanic, 7 % were Asian, African American, Middle Eastern, mixed, or others, and 20 % were unreported. Finally, 38 % were in grade 8, 20 % grade 7, 28 % grade 9, and 14 % unreported.

Keystroke logging was part of the data collection by design, from which the duration and median time-point of the longest 25 IWIs were extracted. We also obtained several summary process features, including the total number of word inserts and total effective writing time for supplementary analyses.

3.3 Data Analyses

For the first research question, we conducted a hierarchical cluster analyses based on the cosine distance between the IWI vectors using complete linkage (Johnson 1967). The cosine distance measure is used to decide whether the clusters are adequately different from one another. The cosine distance is mathematically equivalent as $(1-r)$, where r is the Pearson correlation coefficient (Jones, Oliphant & Peterson 2014). The Cohen's (1968) rules for evaluating the magnitude of Pearson correlation coefficient r are: no to small association (0–0.30), moderate association (0.31–0.50), and large association (0.51–1.00). We transformed these criteria suggested by Cohen (1968) (for evaluating the size of the correlation coefficient) to the cosine distance measure's scale. Hence, we considered a cosine distance value of equal to or greater than 0.50 to indicate a strong separation between clusters, a value in between 0.30 and 0.50 to indicate a moderate separation, and a value of less than 0.30 to indicate a weak separation.

After the number of clusters was determined, we examined and compared the clusters on several aspects. One aspect was the IWI duration and median time-point pattern. For this analysis, we evenly divided the time axis into 10 bins, with each bin accounting for one tenth of the normalized total writing time.[1] Second, we computed the mean of the logarithm-transformed IWI duration values in each bin separately for each cluster. This statistical transformation was undertaken in order to make the distribution of the IWI duration more similar to a normal distribution (Kalbfleisch & Prentice 2002). Third, separately for each of the 20 bins (i.e., 10 bins in essay prompt), we conducted a two-sample independent t-test between the clusters on the mean log IWI values.

A second aspect we used to compare clusters was the density distribution of the IWI duration and median time-point. We generated heat maps of the IWI duration by median time-point to visually compare the probability of IWIs of certain lengths occurring at certain times in the writing process between clusters. To generate the density graphs, we evenly divided the x axis (normalized median time-point ranging from 0 to 1) into 100 steps, and evenly divided the y axis (IWI duration ranging from 0 to 16 s) into 400 steps, which resulted in 100×400 blocks. Note that the observed maximum IWI duration is much greater than 16 s; however, the overwhelming majority of the information falls under the lower part of each density plot. We then calculated the density of each block and produced a normalized density distribution in the form of a heat map for each cluster.

Finally, we compared the clusters based on the human ratings on text production skills, total number of words produced, and total writing time. We computed the effect sizes and conducted two-sample independent t-tests on the means of those measures between the clusters. Of note is that the total number of words produced

[1] On one hand, the bin size needs to be large enough so that there are enough keystrokes in each bin. On the other hand, it needs to be small enough to show variations across bins. After a number of experiments, we found that ten bins are optimal.

is different from the traditional essay length. That is, words that were produced but deleted during the processes (and not appearing in the final product) were counted for this measure. Further, for total writing time, we used active writing time, which excluded pre-writing (i.e., pause before typing the first character), in order to be consistent with the timing data used for IWI analyses. Finally, due to the highly skewed distribution in human response time, we used logarithm-transformed values for these analyses (Ulrich & Miller 1993; van der Linden 2006).

For the second research question, we examined how the IWI information related to human scores. We used leave-one-out multiple linear regression of human scores on the IWI vector (i.e., top 25 IWI durations and median time-points), and computed the correlation coefficient of observed and predicted human scores. We hypothesized a moderate positive association between IWI information and human score because the processes indexed by the location and duration of the IWI should theoretically contribute to the quality of the final product. In addition to using all samples available for each essay task, to investigate whether there was a difference in association between IWI information and human scores between the clusters, we conducted the regression analyses separately for each cluster.

4 Results

4.1 Results for Research Question 1: Writing Patterns

Based on our criteria, we identified two clusters for both essay tasks, as can be seen in Fig. 3. The cosine distance between the two clusters is 0.63 for the policy recommendation essay and 0.62 for the argumentation essay. Based on the transformed Cohen (1968) criteria, this result indicates that the two clusters are

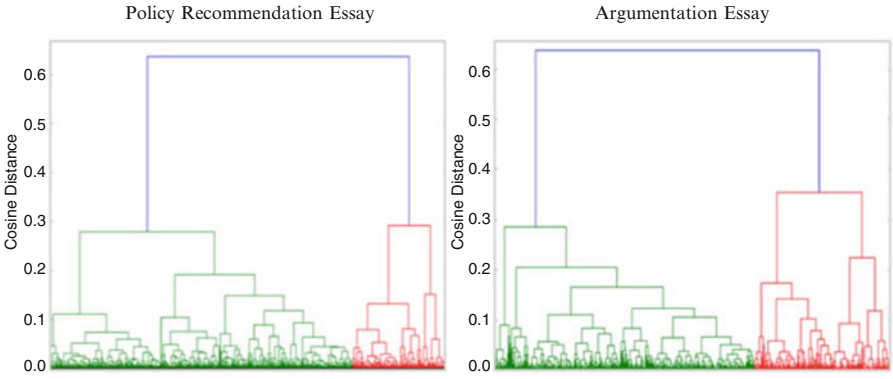


Fig. 3 Clustering of keystroke logs. *Note*: *Green*: Cluster 1. *Red*: Cluster 2. The height of horizontal lines connecting two clusters indicates the cosine distance between two clusters

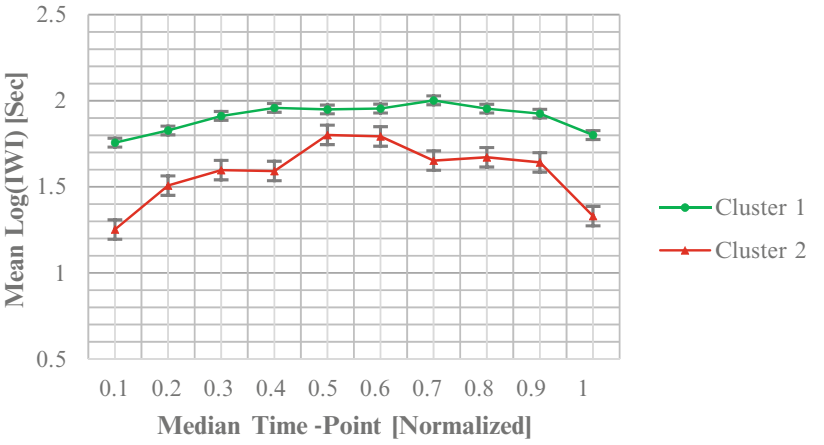


Fig. 4 Mean log(IWI) pattern (policy recommendation essay)

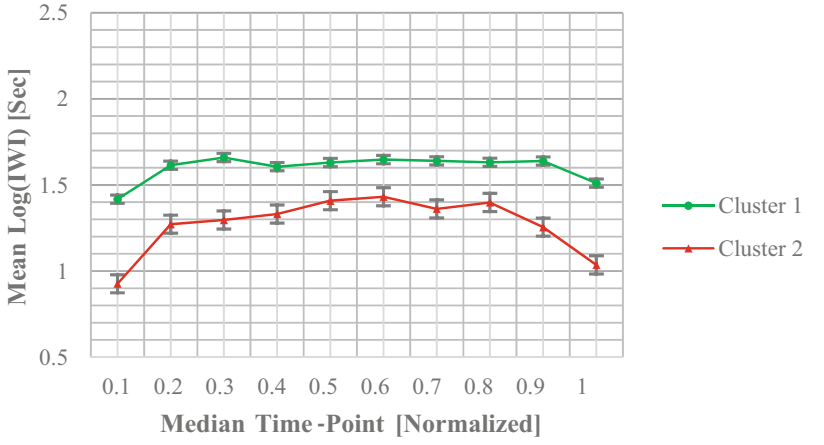


Fig. 5 Mean log(IWI) pattern (argumentation essay)

weakly correlated (i.e., strongly separated). For both essays tasks, there are more essays in cluster 1: for the policy recommendation essay, n = 635 in cluster 1, n = 196 in cluster 2; for the argumentation essay, n = 591 in cluster 1, n = 311 in cluster 2.

Next, we examined the qualitative differences between the two clusters. Figures 4 and 5 show the mean logarithm-transformed IWI duration at different median time-points in the writing process. Based on the t-test results, all pairs of comparisons suggested statistically significant differences (at a $p < 0.05$) between the two clusters. Specifically, for both writing tasks, IWIs in cluster 1 were significantly longer than the ones in cluster 2 consistently throughout the composition. The differences appear to be smaller around the middle than at the two ends of the writing process.

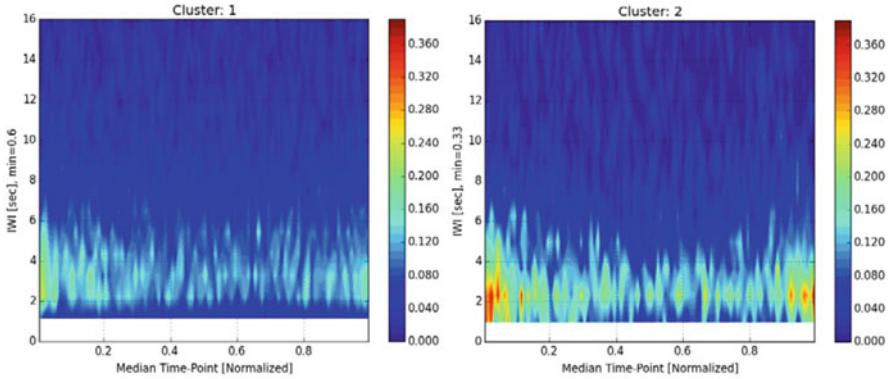


Fig. 6 Visualizing the density of IWI (policy recommendation essay)

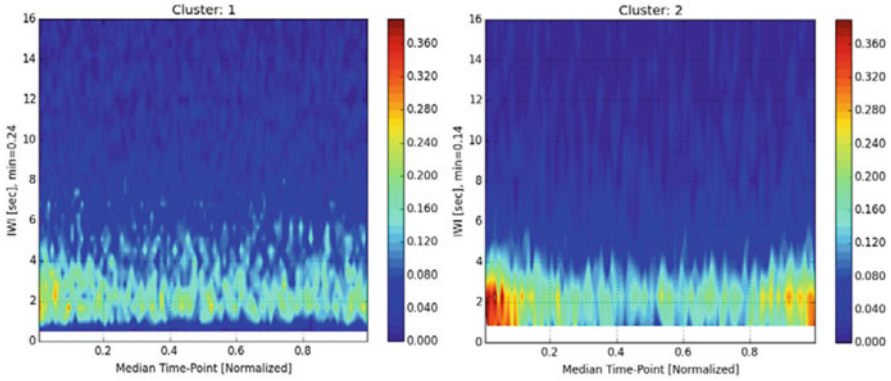


Fig. 7 Visualizing the density of IWI (argumentation essay)

Additionally, it is worth noting that, for both essays, cluster 1 shows a more even straight-line pattern across median time-points than cluster 2, which appears to have a discernable drop in the IWI duration at the beginning and end of the composition process.

To further visually examine how the distribution of the IWIs from the two clusters differs, we plotted the density of IWI and the corresponding median time-points (Figs. 6 and 7). The generation of the density graphs was described previously in the Method section. For cluster 1 on the policy recommendation essay, we found that the density of IWIs appears to hover around 3–4 s throughout the text production process, with density shorter than 2 s fairly low. In contrast, the density pattern in cluster 2 is notably different, where we find higher density with shorter IWIs at the beginning and the end of the composition process.

The contrast in IWI density pattern between clusters 1 and 2 is even more dramatic for the argumentation essay. Cluster 1 shows more evenly distributed IWIs hovering around 2 s throughout the composition, whereas cluster 2 exhibits more

Table 1 Characteristics of essay responses by cluster

Cluster	N	Human score Mean (SD)	Word insert Mean (SD)	Writing time Mean (SD)
Policy recommendation essay				
1	635	2.58 (0.80)	230 (123)	459 (239)
2	196	2.27 (0.81)	181 (111)	382 (247)
Effect size	ES = 0.39	ES = 0.49	ES = 0.37	
t-Value	4.73 ($p < 0.0001$)	5.99 ($p < 0.0001$)	4.59 ($p < 0.0001$)	
Argumentation essay				
1	591	2.87 (0.85)	233 (110)	353 (203)
2	311	2.57 (0.90)	178 (97)	293 (189)
Effect size	ES = 0.35	ES = 0.57	ES = 0.32	
t-Value	4.94 (p < 0.0001)	8.10 (p < 0.0001)	4.54 (p < 0.0001)	

Note: Word insert: sum of words that an author inserts which may or may not appear in the final product. Writing time (in sec.): active composition time excluding pause time before the first keystroke. Effect size and t-value for the writing-time measure are based on log-transformed values

frequent and shorter pauses at the two ends of the writing process. These results of the contrasting IWI patterns between the two clusters also agree with the previous analyses presented in Figs. 4 and 5.

Table 1 shows additional differences between the clusters. For both essays, cluster 1 writers were more able than cluster 2 writers evidenced by the higher human scores on the essay quality. Based on the t-test results, the average human scores for cluster 1 were both statistically significantly higher than cluster 2, and practically significant in terms of effect size. Cluster 1 also shows similarities to previous research results in which students who scored higher spent more time on task, wrote longer essays, and were more fluent in executing critical writing processes (e.g., Deane & Zhang 2015).

4.2 Results for Research Question 2: Relation to Human Ratings

The second research question examines the association of IWI duration and median time-point with human scores; that is, the correlation between information extracted from the writing processes with the quality of the final product. In particular, we conducted regression analyses to predict writing scores from the vectors for the top 25 IWIs, and correlated the resulting predicted scores with human ratings.

Ignoring cluster membership, the correlation coefficient between predicted and observed human scores was 0.46 for the policy recommendation essay (n = 831) and 0.48 for the argumentation essay (n = 902). However, when cluster membership was taken into account, we found notable differences between the clusters.

The cluster-specific regression models resulted in correlation coefficients of 0.47 for cluster 1 and 0.65 for cluster 2 for the policy recommendation essay. Similar results were observed for the argumentation essay, where cluster 1 also yielded a considerably lower correlation coefficient (0.46) than cluster 2 (0.51).

5 Discussion

Our analysis focused on the 25 longest inter-word pauses in each essay and indicated that student response patterns fell into two major groups. In pattern 1, the 25 longest pauses were distributed relatively evenly throughout the composition. Essays that exemplify pattern 1 received higher mean human scores, contained more words, and were composed over a longer active writing time. In pattern 2, the longest 25 pauses were somewhat shorter than in pattern 1, and were concentrated at the beginning and end of the composition. Essays that exemplify pattern 2 received lower mean human scores, had fewer words, and were written over a shorter active composition time. We replicated these findings across two writing prompts, each focused on a different writing purpose and administered to different student samples. It is worth stressing that the results of writing patterns should be interpreted at the group level; that is, the patterns do not reflect the distribution of the 25 longest IWIs of any individual KL.

In the literature, pauses of more than 2 s are generally described as terminating 'bursts' of text production (Chenoweth & Hayes 2001), and tend to be interpreted in think-aloud protocols as occasions for sentence-level planning (Baaijen et al. 2012). This cognitive interpretation can readily be applied to pattern 1. As can be observed in Figs. 6 and 7, the longest pauses produced by pattern-1 writers fell most often between 1.5 and 4 s in duration, typical of pauses between bursts. This interpretation is strengthened by the fact that pattern-1 pauses were evenly distributed across the entire text. The resulting rhythm—a regular series of bursts of fast production delimited by pauses of 2 s or more—may be emblematic of fluent text production, in which the writer pauses primarily to plan the next major grammatical or textual unit.

The striking feature of pattern 2 is the presence of a second kind of pause, mostly shorter than the pauses observed in pattern 1, concentrated near the beginning and the end of the composition. These time points are arguably where a writer who has difficulty generating text is most likely to experience difficulty, consistent with Baaijen et al. (2012)'s observation that certain kinds of behavioral events, such as text production followed by revision, are associated with shorter pauses. It is thus possible, though by no means certain, that the higher frequency of short pauses concentrated at the beginning and ends of pattern 2 essays reflects difficulties in text generation, leading to false starts and interruptions instead of fluent text production at the beginning of an essay (when the writer is under the most stress to plan content), and at the end of an essay (when the writer may be running out of ideas, and thus once more experiencing higher levels of uncertainty about what to write next).

We conducted a post-hoc qualitative analysis of a small subset of logs from this dataset, and found that some weaker writers did produce a small amount of text—a few words, or even part of a word—and then delete it after a short pause, only to proceed to another false start. It is thus possible that pattern 2 involves this kind of hesitation, although we cannot confirm it without further analysis in which we correlate the distribution of IWIs with the distribution of deletions and edits.

6 Conclusion

In this study, we propose a new way to compare the temporal sequence of IWIs across different students using a vector representation. This approach enables us to describe global patterns in pausing behavior, which may correspond to different cognitive strategies or styles of writing. This study represents an initial attempt, using a specific keystroke log feature (IWIs) and a specific similarity metric, to explore ways to represent and directly compare KLs, analyze the resulting classification patterns, and pose cognitive accounts for the identified patterns in the context of writing done for standardized tests. Overall, our analysis indicates that there do appear to be qualitative differences among groups of writers in the time-course of text production, some of which differences can be detected from a very small sample of events (e.g., only the 25 longest inter-word intervals).

However, it should be note that the method we employed in this study represents our starting point to explore better representations and similarity measures for the KLs. Based on the current methodological scheme, we observed some clear pattern differences in students' writing processes, which held across two prompts. However, a different scale transformation, for example, on the IWI time-point, will change the similarity matrix structure and affect the clustering results. In our future investigations, we will experiment with other similarity measures (e.g., Euclidean or Mahalanobis types of distance measures) and representations such as matched filtering, which might be more robust than the current approach.

It is also important to note that the decision to target the 25 longest IWIs represents two levels of abstraction: first, by restricting attention to IWIs, and second, by excluding shorter IWIs from the analysis. These decisions provided a useful lens with which to examine the data, since the literature provides strong reasons to suspect that the longest IWIs will reflect global differences in writing patterns and strategies. The decision to standardize to the 25 longest IWIs also made it easier to compare essays of different lengths (and which were composed over shorter or longer time periods), but it does represent a small portion of the total data; hence, it will be useful to extend the scope of future analysis to include all IWIs.

Deane (2014) provides evidence that many keystroke features are not particularly stable across changes in prompt, genre, and/or topic. Therefore, caution should be exercised in generalizing the results. Further studies are needed to determine the extent to which these results reflect prompt-specific or general differences in student

writing behaviors, which will require studying students from other grade levels and writing prompts targeting other writing purposes (e.g., narrative writing).

Finally, it might be valuable to enrich the representation to include information about the context of such writing actions as IWI. For example, some IWIs happen between words in a long burst of text production; others, in the context of other actions, such as edits or deletions. We would interpret the second cluster, in which most pauses were near the beginning and end of student essays, very differently if they were associated with editing and deletion, than we would if they were associated with uninterrupted text production. Thus, it would be of particular value to enrich the current approach by undertaking analyses that identify qualitative, linguistic or behavioral differences and that would allow us to relate those findings to the differences in writing patterns observed here.

Acknowledgements We would like to thank Marie Wiberg, Don Powers, Gary Feng, Tanner Jackson, and Andre Rupp for their technical and editorial suggestions for this manuscript, thank Randy Bennett for his support of the study, and thank Shelby Haberman for his advice on the statistical analyses in this study.

References

Almond, R., Deane, P., Quinlan, T., & Wagner, M. (2012). *A preliminary analysis of keystroke log data from a timed writing task (RR-12-23)*. Princeton, NJ: ETS Research Report.

Alves, R. A., Castro, S. L., & de Sousa, L. (2007). Influence of typing skill on pause–execution cycles in written composition. In G. Rijlaarsdam (Series Ed.), M. Torrance, L. van Waes, & D. Galbraith (Vol. Eds.), *Writing and cognition: Research and applications* (Studies in Writing, Vol. 20, pp. 55–65). Amsterdam: Elsevier.

Baaijen, V. M., Galbraith, D., & de Glopper, K. (2012). Keystroke analysis: Reflections on procedures and measures. *Written Communications, 29*, 246–277.

Banerjee, R., Feng, S., Kang, J. S., & Choi, Y. (2014). *Keystroke patterns as prosody in digital writings: A case study with deceptive reviews and essays*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar.

Beauvais, C., Olive, T., & Passerault, J. (2011). Why are some texts good and others not? Relationship between text quality and management of the writing processes. *Journal of Educational Psychology, 103*, 415–428.

Bennett, R. E. (2010). Cognitively Based Assessment of, for, and as Learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement, 8*, 70–91.

Bennett, R. E., Deane, P., van Rijn, P. (2016). From cognitive-domain theory to assessment practice. Educational Psychologist, 51, 82–107.

Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in writing: Generating text in L1 and L2. *Written Communication, 18*, 80–98.

Chukharev-Hudilainen, E. (2014). Pauses in spontaneous written communication: A keystroke logging study. *Journal of Writing Research, 6*, 61–84.

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

Deane, P. (2014). *Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks (RR-14-03)*. Princeton, NJ: ETS Research Report.

Deane, P., Sabatini, J. S., Feng, G., Sparks, J., Song, Y., Fowles, M., et al. (2015). *Key practices in the English Language Arts (ELA): Linking learning theory, assessment, and instruction (RR-15-17)*. Princeton, NJ: ETS Research Report.

Deane, P., & Zhang, M. (2015). *Exploring the feasibility of using writing process features to assess text production skills (RR-15-26)*. Princeton, NJ: ETS Research Report.

Dragsted, B., & Carl, M. (2013). Towards a classification of translation styles based on eye-tracking and keylogging data. *Journal of Writing Research, 5*, 133–158.

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

Gould, J. D. (1980). Experiments on composing letters: Some facts, some myths, and some observations. In L. Gregg & E. Steinberg (Eds.), *Cognitive processes in writing* (pp. 97–127). Hillsdale, NJ: Lawrence Erlbaum.

Grabowski, J. (2008). The internal structure of university students' keyboard skills. *Journal of Writing Research, 1*, 27–52.

Hao, J., Smith, L., Mislevy, R., von Davier, A., & Bauer, M. (2016). *Taming log files from game and simulation-based assessment: Data model and data analysis tool*. (RR-16-10) Princeton, NJ: ETS Research Report.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*, 241–254.

Jones, E., Oliphant, T., & Peterson, P. (2014). *SciPy: Open source scientific tools for Python* [Computer software]. Retrieved from http://www.scipy.org/.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd ed.). Hoboken, NJ: Wiley.

Leijten, M., Macken, L., Hoste, V., van Horenbeeck, E., & van Waes, L. (2012). *From character to word level: Enabling the linguistic analyses of Inputlog process data*. Proceedings of the EACL 2012 Workshop on Computational Linguistics and Writing, Avignon, France.

Leijten, M., & van Waes, L. (2013). Keystroke logging in writing research using Inputlog to analyze and visualize writing processes. *Written Communication, 30*, 358–392.

Leijten, M., van Waes, L., Schriver, K., & Hayes, J. R. (2014). Writing in the workplace: Constructing documents using multiple digital sources. *Journal of Writing Research, 5*, 285–377.

Miller, K. S. (2000). Academic writers on-line: Investigating pausing in the production of text. *Language Teaching Research, 4*, 123–148.

Roca de Larios, J., Manchon, R., Murphy, L., & Marin, J. (2008). The foreign language writer's strategic behavior in the allocation of time to writing processes. *Journal of Second Language Writing, 17*, 30–47.

Ulrich, R., & Miller, J. (1993). Information processing models generating log normally distributed reaction times. *Journal of Mathematical Psychology, 37*, 513–525.

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

van Waes, L., Leijten, M., & van Weijen, D. (2009). Keystroke logging in writing research: Observing writing processes with Inputlog. GFI-Journal, No 2-3.

Xu, X., & Ding, Y. (2014). An exploratory study of pauses in computer-assisted EFL writing. *Language Learning & Technology, 18*, 80–96.

Zhang, M., & Deane, P. (2015). *Process features in writing: Internal structure and incremental value over product features (RR-15-27)*. Princeton, NJ: ETS Research Report.

Identifying Useful Features to Detect Off-Topic Essays in Automated Scoring Without Using Topic-Specific Training Essays

Jing Chen and Mo Zhang

Abstract *E-rater*® is the automated scoring engine used at ETS to score the writing quality of essays. A pre-screening filtering system is embedded in *e-rater* to detect and exclude essays that are not suitable to be scored by *e-rater*. The pre-screening filtering system is composed of a set of advisory flags, each of which marks some unusualness of the essay (e.g. repetition of words and sentences, restatement of the prompt). This study examined the effectiveness of an advisory flag in the filtering system that detected off-topic essays. The detection of off-topic essays usually requires topic specific training essays to train the engine in order to identify essays that are very different from the other essays of the same topic. The advisory flag used here is designed to detect off-topic essays without using topic-specific training essays because topic-specific training essays may not available in real test settings. To enhance the capability of this off-topic advisory flag, we identified a set of essay features that are potentially useful in distinguishing off-topic essays that do not require topic specific training essays. These features include essay length, the number of word types (exclude non-content-bearing words), the number of word tokens, the similarity of an essay to training essays, essay organization, and the variety of sentences in an essay.

Keywords Automated essay scoring • Feature selection • Off-topic essay detection

1 Introduction

Automated scoring is now more and more widely used to score constructed response items given its advantages such as low cost, real-time feedback, quick score-turnaround and consistency over time (Williamson, Bejar & Hone 1999). Automated scoring engines have been developed to score different types of constructed response such as short responses, essays and speaking responses. Automated Essay Scoring

J. Chen (✉) • M. Zhang
Educational Testing Service, Princeton, NJ 08541, USA
e-mail: jingchenhao@gmail.com

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer
Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_24

(AES), which is the subject of this paper, is defined as using computer technology to evaluate and score the written prose (Dikli 2006). The scoring process of an AES system usually involves extracting features using Natural Language Processing (NLP) techniques and statistical modeling that predict human scores based on the extracted essay features. Though the performance of AES systems may equal or surpass that of human raters in many aspects of writing, the systems do not really "read" or "understand" the essays. If an essay is very unusual, an AES system may fail to process the essay or it may assign a score that does not reflect the criteria specified in the scoring rubrics.

1.1 The Filtering System of Automated Essay Scoring

A pre-screening filtering system is often used to identify unusual essays that are not suitable to be scored by the automated scoring system. An effective pre-screening filtering system is important to ensure the validity of automated scores. If the scoring engine handles problematic responses in the same way as it handles the other normal responses, it may degrade users' confidence in using the scoring engine. Furthermore, essays detected as unusual by the filtering system often need to be scored by human raters. An effective filtering system would detect unscorable essays while minimizing the number of essays that are falsely identified as unscorable to avoid unnecessary human scoring cost.

Several widely used AES systems have pre-screening filtering systems to detect unusual responses. The Intelligent Essay Assessor (IEA, Landauer, Laham & Foltz 2003) checks for a number of things such as the number and degree of clustering of word type repetitions, the extent to which the essay is off topic and whether the essay is a copy or rearrangement of another essay (Wollack & Fremer 2013). The IntelliMetric system developed by Vantage Learning (Elliot 2003; Shermis & Barrera 2002) has warning flags for things such as nonsensical writing, violent language, copying the prompt and plagiarism (Rudner, Garcia & Welch 2006). *E-rater*®, the automated scoring engine developed at ETS (Attali & Burstein 2006) has a set of advisory flags to identify atypicality in an essay (e.g., too many repetitions of words and sentences, unusual organization, off-topic content).

This study investigates the effectiveness of an *e-rater* advisory flag that is designed to detect off-topic essays. In high-stakes writing assessments, if an essay triggers any advisory flag, the essay will be excluded from automated scoring and scored by human raters only. Thus, an effective advisory flag would detect unsuitable responses while minimizing the number of essays that are falsely identified as unsuitable for automated scoring in order to save human scoring cost. For low-stakes writing assessments, e-rater is often used as the sole scoring method. If an essay triggers the off-topic advisory flags, e-rater will still generate a score for the essay but provides a warning to indicate the essay might be off-topic so the score assigned may not be valid. The assessments used in this study are all high-stakes assessments which means essays that triggered the off-topic advisory flag will be

excluded from automated scoring and scored by human raters only. We examine the performance of an off-topic advisory flag and identify potentially useful features to detect off-topic essays more accurately. In the following section, we introduce our definition of off-topic essays and the off-topic advisory flag we evaluated in this study.

1.2 Off-Topic Essays and the E-rater® Off-Topic Advisory Flags

Off-topic essays are often off-topic in many divergent ways. We define off-topic essays somewhat broadly to include unexpected topic essays, bad faith essays, essays in a foreign language, essays consisting of only keystroke characters with little lexical content or essays that are illegible. In scoring rubrics, these essays usually receive human scores of '0'. Two common types of off-topic essays in writing assessments are essays written on an unexpected-topic and bad faith essays. An essay of unexpected-topic can be well-written, but it provides no evidence of an attempt to address the assigned topic. The bad faith essays are usually written by examinees who respond uncooperatively. They write text irrelevant to the assigned topic because of boredom or other reasons.

In the pre-filtering system of *e-rater*, there are two types of advisory flags that detect off-topic essays. One type of off-topic advisory flags require topic specific training essays to train the advisory in order to identify essays that are very different from the other essays on the same topic. The other type of off-topic advisory flag does not require topic specific training essays. The detection of off-topic essays relies on essay features that can be computed without topic specific training essays. In real test settings, topic-specific training essays may not available when new prompts are administered. In addition, the sample size of topic-specific training essays may not be sufficient sometimes. An advisory flag that does not need topic-specific training essays can be used more flexibly and broadly in real test settings. Thus, we only investigate this type of off-topic advisory flag in this paper and try to find essay features that can be computed without topic-specific training essays to improve the detection of off-topic essays.

More specifically, the advisory flag we evaluate in this study detects off-topic essays based on the similarity between the text of an essay and the prompt on which the essay is supposed to have been written (Higgins, Burstein, & Attali, 2006). This similarity is measured by a feature (abbreviated as "S_Prompt") calculated based on Content Vector Analysis (CVA, for a detailed introduction, see Kaplan 2010, p. 531). The similarities between an essay and its target prompt (i.e. S_Prompt) as well as the reference prompts are calculated and sorted. If the similarity between an essay and its target prompt ranked amongst the top 15 % of the similarity scores, then the essay is considered on topic. Otherwise, it is identified as off-topic.

This study evaluates the effectiveness of the off-topic advisory flag introduced above. Besides the feature used in this flag, e-rater extracts a lot of other essay

features to evaluate essays. This study explores what additional essay features can potentially be used to detect off-topic essays. The study is guided by the following two research questions:

1. How effective is the advisory flag in detecting off-topic essays?
2. Among the essay features that *e-rater* extracts, what are the potentially useful ones that can detect off-topic essays?

2 Methods

2.1 Data

The data for this study came from the writing tasks of two large-scale high-stakes assessments. Assessment I is a college level test, and Assessment II is an English proficiency test. The writing section in Assessment I includes two tasks, which we refer to as Task A and Task B for the purpose of this paper. Task A requires examinees to critique an argument while Task B requires examinees to articulate an opinion and support their opinions by using examples or relevant reasoning. Similar to the writing tasks of Assessment I, the writing section of Assessment II also included two tasks, which we refer to as Task C and Task D. Task C requires test takers to respond in writing by synthesizing the information that they had read with the information they had heard. Task D requires test takers to articulate and support an opinion on a topic.

The score scale of Task A and B is from 1 to 6, and that of Task C and D is from 1 to 5. The lowest score, 1, indicates a poorly written essay and the highest score, 5 or 6, indicates a very well written essay. Specifically, the scoring rubrics of these four writing tasks all specify that an essay at score level '0' is not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank. Therefore, for the purpose of this study, we classify all the essays that received a human score of '0' as off-topic essays (except the blank ones) and the other essays with non-zero scores as on-topic essays.

In operational scoring design, essays from high-stakes assessments usually are scored by a human rater first. If the human rater assigns a score of '0' to an essay which indicates that the essay is very unusual, the essay will be excluded from automated scoring entirely. Instead, a second human rater will evaluate this essay to check the score from the first human rater. Because off-topic essays will be flagged by human raters, the issue of off-topic responses is not viewed as a serious problem for automated scoring in high-stakes assessments. However, in low-stakes assessment when e-rater is used as the primary or sole scoring system, it's important to have an effective flag to detect off-topic essays that may not suitable for automated scoring.

To evaluate the effectiveness of the off-topic flag discussed previously, we selected a random sample of around 200,000 essays from each writing task that was

Table 1 Essay sample size of each writing task (Sample 1)

	Task A	Task B	Task C	Task D
No. of selected essays	199,650	199,656	200,782	199,605
Proportion of off-topic essays (%)	0.02	0.03	0.7	0.08

Table 2 Number of selected off-topic and on-topic essays (Sample 2)

	Off-topic	On-topic
Task A	388	388
Task B	809	809
Task C	24,244	24,244
Task D	3147	3147

collected during July 2012 to June 2013 for Assessment I and during July 2011 to June 2013 for Assessment II. These random samples include both off-topic and on-topic essays. Table 1 lists the precise number of essays selected from each writing task and the proportion of off-topic essays in each sample.

In operational scoring, an essay will be scored by two human raters if the first human rater assigned a score "0". Our off-topic essay sample only includes essays that received score "0" from both human raters to ensure that judgment of score "0" is agreed by both human raters. Among the essays that received score "0" from both human raters, we also excluded essays that did not meet the length requirement of containing at least two sentences. These extremely short essays will be excluded from automated scoring by an advisory flag that detects extremely short essays. Thus, we do not need to consider how well the off-topic advisory flag detects these extremely short off-topic essays. Because of these reasons, the proportion of off-topic essays listed in Table 1 is lower than the proportion of essays that received score "0" in operational scoring.

The sample listed in Table 1 (Sample 1) is used to evaluate the effectiveness of the off-topic advisory flag (e.g. calculate precision and recall rates). In this sample, there are only small numbers of off-topic essays from Task A and Task B (i.e. fewer than 100 essays). To compare the linguistic features of on-topic and off-topic essays and identify the most distinctive features, we selected off-topic essays from a broader time range to have more off-topic essays. For each writing task, we selected all the off-topic essays (i.e. essays that received score "0" from both human raters and longer than two sentences) from operational scoring during Jul. 2011 to Jun. 2015. We randomly selected a set of on-topic essays (i.e. essays that received non-zero scores from human raters) from each writing task from the same time range to match the sample size of the selected off-topic essays. Table 2 presents the resulting sample sizes of these off-topic and on-topic essays. This is our second sample. For both groups of essays, we extract all the essay features using the latest version of the *e-rater* engine.

In our analysis, we include nine high-level features that predict human scores, their associated low-level features and some additional features that are not used in predicting human scores but are used to provide additional information about the

essays. The nine high-level features are grammar, usage, mechanics, development, organization, word choice, word length, collocation and preposition, and sentence variety. Most of these nine features are composed of sets of low-level features computed using Natural Language Processing (NLP) techniques that are then combined to produce the high-level feature values. The features used to provide additional information about the essays included features that measure essay length (e.g. number of sentences), features related to word type and word token usage and features that measure the similarity between an unseen essay and the training essays.

More specifically, among the features that provide additional information, one feature is the number of word types, which is a count of the number of unique words in the essay. Another feature is the number of word token, which is the frequency of unique words. If a unique word type appears multiple times in an essay, "the number of word types" will only count once, but "the number of word tokens" will count multiple times. When calculating "the number of word types" and "the number of word tokens" features, a "stop list" is used to exclude non-content-bearing words (e.g. words such as "the", "of") from the calculation. So these two features only count unique content-bearing words. Another feature, ZTT (Z-score of the ratio of the number of word types to the number of word tokens), provides a measure of the variety of words in an essay. The feature value will be high if each unique word only appears once and will be low if each unique word appears many times.

The similarity between an essay and the training essays is measured by several features including S_Max, S1, S2, S3, S4, S5 and S6. A similarity value between an unseen essay and each of the training essays can be calculated. S_Max is the largest similarity value among all these similarity values. The S1 feature measures the similarity between an unseen essay and all the training essays that received score "1" assigned by human raters. Similarly, feature S2 measures the similarity between an unseen essay and all the training essays that received score "2" and so on. All these similarity features are calculated based on CVA.[1]

2.2 Data Analysis

First, to find out the effectiveness of the off-topic advisory flag, we evaluated the precision, recall and F-score of this flag in detecting off-topic essays based on sample 1. Precision is the proportion of detected off-topic essays that are truly off-topic. Recall is the proportion of true off-topic essays (as classified by human raters) that are detected by advisory flag. F-score[2] is a measure that balances precision and

[1]The essays used to train the CVA based features were collected from Assessment I and II during July 2012 to July 2013 and during July 2011 to June 2013 respectively. Around 100,000 essays were used as training essays for each writing task of the assessments. There were no essays in common between the dataset described in Table 2 and the data used to train these CVA features.

[2]F-score is defined as F = 2* (Precision*Recall)/(Precision + Recall).

recall rates. Second, to identify the features that are potentially useful in detecting off-topic essays, we compare the values of each of the e-rater features between the on-topic and the off-topic essay groups based on sample 2. We calculated the mean and standard deviation of each feature for the two groups of essays and used Cohen's d to measure the difference between the means of these two groups.

For Cohen's d, an effect size of 0.2–0.3 indicates a "small" effect. An effect size around 0.5 suggests a "medium" effect and an effect size of 0.8 to infinity suggests a "large" effect (Cohen 1988). Features with large Cohen's d (greater than 0.8) are considered as the most distinctive features. We examined whether these identified features can be grouped into some common themes that measure particular aspects of writing.

3 Results

3.1 Effectiveness of the Off-Topic Advisory Flag

Table 3 presents the precision, recall and F-score of the advisory flag we evaluated. The precision rate is 100 % across four writing tasks. All the essays detected as off-topic are classified as off-topic by human raters as well. The high precision rate will save human scoring cost associated with scoring false positive cases. The recall rate varies between 2.2 and 18.1 % across four writing tasks. Improvements could be made to capture more off-topic essays. Therefore, we investigated what additional essay features can potentially be used to detect off-topic essays.

3.2 Most Distinctive Features Between On-Topic and Off-Topic Essays

Features with large Cohen's d (greater than 0.8) are identified as the most distinctive features for off-topic essays. These features can be grouped into five categories, each of which measures a particular aspect of writing. In this section, we describe the difference between the on-topic and the off-topic essays in these five aspects.

Essay length. Off-topic essays are considerably shorter than on-topic essays. Table 4 presents some descriptive statistics of essay length of the on-topic and the off-topic essays from each writing task. Essay length is calculated from the number

Table 3 The precision, recall and f-score of the off-topic advisory flag

	Task A	Task B	Task C	Task D
Precision	100.0	100.0	100.0	100.0
Recall	16.7	16.0	2.2	18.1
F-score	28.6	27.6	4.4	30.7

Table 4 Essay length of the on-topic and off-topic essays

		On-topic		Off-topic		
		Mean	SD	Mean	SD	Cohen's d
Task A	Chars	1961.08	645.09	945.01	718.54	1.49
	Words	395.61	127.53	185.58	138.87	1.58
	Sentences	18.17	5.97	9.01	6.72	1.44
Task B	Chars	1987.26	654.77	1244.92	696.06	1.10
	Words	406.73	133.38	264.52	143.54	1.03
	Sentences	19.25	6.68	13.81	7.66	0.76
Task C	Chars	1068.53	272.14	727.59	327.42	1.13
	Words	216.30	54.56	141.95	63.40	1.26
	Sentences	11.27	3.48	8.54	3.77	0.75
Task D	Chars	1597.08	416.08	765.39	550.88	1.70
	Words	338.54	85.37	167.02	121.69	1.63
	Sentences	18.00	5.65	11.45	8.89	0.88

Table 5 Number of word types and number of word tokens of the on-topic and off-topic essays

		On-topic		Off-topic		
		Mean	SD	Mean	SD	Cohen's d
Task A	Number of word types	118.11	33.47	59.07	28.96	1.89
	Number of word tokens	211.50	68.56	113.80	84.74	1.27
	ZTT	0.00	0.93	0.71	2.25	−0.41
Task B	Number of word types	131.13	41.16	83.06	39.59	1.19
	Number of word tokens	221.45	73.18	151.34	89.19	0.86
	ZTT	−0.04	0.99	−0.17	1.60	0.10
Task C	Number of word types	78.12	17.35	62.69	21.48	0.79
	Number of word tokens	117.89	29.24	86.47	38.50	0.92
	ZTT	0.03	0.97	0.97	1.16	−0.88
Task D	Number of word types	113.03	29.28	60.62	39.15	1.52
	Number of word tokens	190.44	48.60	103.15	77.61	1.35
	ZTT	0.06	0.96	0.89	2.17	−0.49

of characters, words, and sentences. Across all four writing tasks and all three measures of essay length, off-topic essays are considerably shorter than on-topic essays. For example, on average, the on-topic essays from Task A have around 395 words. However, the off-topic essays from the same task only have around 186 words.

The number of word types and the number of word tokens. Our analysis reveals that the off-topic and on-topic essays show a large difference in terms of two features: the number of word types and the number of word tokens. Table 5 presents the statistics of these two features, which shows that off-topic essays have fewer unique words and lower occurrences of unique words compared to on-topic essays.

However, these two features may closely related to essay length since longer essays by definition have more words (new or repeated). So we looked at another feature that measures the word variety of an essay but has less influence from essay length. This feature (ZTT) is based on the Z-score of the ratio of the number of word types to the number of word tokens. Table 5 lists descriptive statistics of this feature. In general, ZTT of off-topic essays are higher than that of on-topic essays except Task B. This pattern is reasonable because when an essay has many unique words without any repetition, the essay may fail to focus on key concepts and stay on-topic. However, future research is needed to examine whether ZTT is really effective in predicting off-topic essays across different types of writing tasks since the pattern from Task B is not consistent with the others.

Similarity features. Table 6 lists the mean, standard deviation of the similarity features for the off-topic and on-topic essay groups and the Cohen's d of the mean difference. The similarity between an off-topic essay and the training essays is much lower than that between an on-topic essay and the training essays. Some off-topic essays, such as bad faith essays or essays written in a foreign language are expected to have very low similarity to the majority of essays. So it is reasonable that on average, the similarity features of off-topic essays are lower than those of on-topic essays. Though the size of mean difference varies across four tasks, in general, the similarity features of the off-topic essays are much lower than those of the on-topic essays. These similarity features can potentially be used to distinguish off-topic essays from the on-topic ones.

Organization. The organization feature score of off-topic essays is significantly lower than that of on-topic essays. The organization feature consists of a set of low-level features that detect whether particular discourse elements are present or absent in an essay (Burstein, Marcu & Knight 2003). These particular discourse elements include introductory material (to provide the context or set the stage), a thesis statement (to state the writer's position in relation to the prompt), main ideas (to assert the author's main message), supporting ideas (to provide evidence and support the claims in the main ideas, thesis, or conclusion), and a conclusion (to summarize the essay's entire argument) (Attali & Burstein 2006).

Table 7 lists the descriptive statistics of the organization feature. The organization feature of the off-topic essays is considerably lower than those of the on-topic essays across all four writing tasks. Some bad faith essays might not be argumentative or summary-like in nature, which might lead to the lack of organizational elements that are typical of argumentative writing (e.g. main idea, supporting evidence). Some off-topic essays may not have good organization because they are too short. The difference in the organization feature between on-topic and off-topic essays might reflect the difference in essay length. Thus, the organization feature that detects whether particular discourse elements are present is a potentially useful feature and it needs to be examined to see whether it can provide useful information in addition to essay length in detecting off-topic essays.

Sentence variety. Another finding is that the sentence variety feature values of the off-topic essays are much lower than those of the on-topic essays. The sentence

Table 6 Similarity features of the on-topic and the off-topic essays

		On-topic		Off-topic		
		Mean	SD	Mean	SD	Cohen's d
Task A	S_Max	0.11	1.02	−0.37	1.76	0.33
	S1	0.07	0.04	0.08	0.06	−0.20
	S2	0.16	0.04	0.12	0.04	1.00
	S3	0.18	0.04	0.12	0.05	1.33
	S4	0.18	0.04	0.11	0.05	1.55
	S5	0.16	0.04	0.11	0.05	1.10
	S6	0.10	0.07	0.07	0.06	0.46
Task B	S_Max	0.07	0.97	−1.16	1.28	1.08
	S1	0.11	0.05	0.08	0.03	0.73
	S2	0.18	0.05	0.13	0.05	1.00
	S3	0.20	0.05	0.13	0.05	1.40
	S4	0.20	0.06	0.13	0.05	1.27
	S5	0.18	0.05	0.11	0.04	1.55
	S6	0.12	0.04	0.07	0.03	1.41
Task C	S_Max	0.01	1.02	−1.23	2.08	0.76
	S1	0.13	0.04	0.09	0.04	1.00
	S2	0.14	0.04	0.09	0.03	1.41
	S3	0.15	0.04	0.09	0.03	1.70
	S4	0.14	0.04	0.09	0.03	1.41
	S5	0.13	0.04	0.08	0.03	1.41
Task D	S_Max	−0.08	0.96	−1.17	1.07	1.07
	S1	0.10	0.03	0.06	0.03	1.33
	S2	0.18	0.05	0.10	0.05	1.60
	S3	0.20	0.05	0.10	0.05	2.00
	S4	0.19	0.04	0.10	0.05	1.99
	S5	0.17	0.05	0.08	0.05	1.80

Table 7 Organization feature of the on-topic and the off-topic essays

	On-topic		Off-topic		
	Mean	SD	Mean	SD	Cohen's d
Task A	1.97	0.38	1.13	0.62	1.63
Task B	1.96	0.34	1.52	0.63	0.87
Task C	1.75	0.42	1.26	0.56	0.99
Task D	1.94	0.32	1.18	0.65	1.48

variety feature is composed of a set of low-level features that measure the occurrence of particular types of words, phrases, and punctuations. A higher sentence variety score indicates that an essay has heterogeneous sentences. The statistics list in Table 8 suggest that off-topic essays have much lower sentence variety scores than on-topic essays. Some examinees could have written homogeneous sentences because of low language abilities. However, it is also possible that the lower sentence variety scores of off-topic essays are due to the fact that off-topic essays are often too short to include a large variety of syntactic types. Thus, sentence variety

Table 8 Sentence variety feature of the on-topic and off-topic essays

	On-topic		Off-topic		
	Mean	SD	Mean	SD	Cohen's d
Task A	3.86	0.45	3.00	0.70	1.46
Task B	3.82	0.45	3.00	0.70	1.39
Task C	3.12	0.48	2.93	0.60	0.35
Task D	3.44	0.45	2.46	0.79	1.52

is a potentially useful feature and future research needs to be done to investigate whether this feature provides useful information in addition to essay length to detect off-topic essays.

4 Discussion

This study investigates the effectiveness of an advisory flag in detecting off-topic essays in automated scoring. A well-formed and well-written essay that does not address the assigned topic may receive an overestimated score from an AES system because of its linguistic features. Successful identification of off-topic essays is essential to ensure the validity of machine scores and to support automated scoring as the primary scoring method.

Our investigation of the effectiveness of the existing advisory flag reveals that this flag has a 100 % precision rate in detecting off-topic essays across the four data sets we evaluated. The recall rate varies around 2.2–18.1 % across four data sets. These results suggest that all detected essays are truly off-topic but a large number of truly off-topic essays are not captured by the advisory flag. To improve the performance of the existing advisory flag, we identified some features that can potentially be used to build new advisory flags to detect more off-topic essays. These features include essay length, the number of word types (excluding non-content-bearing words), the number of word tokens, the word variety feature (ZTT), the similarity of an unseen essay to the training essays, essay organization, and sentence variety.

Two limitations of this study should be noted. First, our evaluation of the performance of the off-topic advisory flag is relatively imprecise. We did not further classify all the essays that received a human score of 0 into different categories according to the way in which an essay diverges from the requested essay topic (e.g. unexpected topic essays, bad faith essays). We lack information to evaluate the performance of the flag in detecting different types of off-topic essays. Second, we only used Cohen's d to identify the features that are potentially useful in detecting off-topic essays. A lot of methods are available for feature selection. For example, Guyon and Elisseeff (2003) introduced variable and feature selection methods such as variable ranking. In future studies, we will apply other feature selection methods and compare the results across methods to provide a better selection of features.

Future research could start with the identified features to build new flagging mechanisms. These identified features can be combined and refined to find out the most effective ones in predicting off-topic essays. The off-topic flag we evaluated only uses the similarity between essay text and prompt text to detect off-topic essays. When an essay triggers the flag, it's easy to tell why the flag is triggered and what kind of problem the essay may have. However, considering one flagging criterion at a time and using a pre-specified triggering threshold may not work as well as building statistical models that use multiple criteria simultaneously and learning from real data to predict the probability of being off-topic. For example, a logistic regression model can be built to predict the likelihood that an essay might be off-topic using features such as essay length, organization, and sentence variety as independent variables.

Finally, additional features such as response time and process data (e.g. essay keystroke) can be collected to predict off-topic essays. For example, if an examinee submits an essay in a very short time after the assessment begins (e.g. 30 s), the essay is likely to be a bad faith essay. Since off-topic essays can be off-topic in many different ways, more features will capture the unusualness of the essays from different aspects, which will help to detect off-topic essays more accurately.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *Journal of Technology, Learning, and Assessment, 4*(3), 1–31. Retrieved from http://www.jtla.org.

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems, 18*(1), 32–39.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning and Assessment, 5*(1), 1–36. Retrieved from http://www.jtla.org.

Elliot, S. (2003). IntelliMetric: From here to validity. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross disciplinary approach*. Mahwah, NJ: Lawrence Erlbaum.

Guyon, I., & Elisseeff, A. (2003). An introduction of variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering, 12*(2):145–159.

Kaplan, R. B. (2010). *The Oxford handbook of applied linguistics*. Oxford, UK: Oxford University Press.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated essay scoring: A cross disciplinary perspective. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring and annotation of essays with the intelligent essay assessor* (pp. 87–112). Mahwah, NJ: Lawrence Erlbaum.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of the intellimetric essay scoring system. *Journal of Technology, Learning and Assessment, 4*(4). Retrieved from http://www.jtla.org.

Shermis, M., & Barrera, F. (2002). *Exit assessments: Evaluating writing ability through Automated Essay Scoring*. ERIC document reproduction service no ED 464 950.

Williamson, D. M., Bejar, I. I., & Hone, A. S. (1999). Mental model comparison of automate and human scoring. *Journal of Educational Measurement, 36*, 158–184.

Wollack, J. A., & Fremer, J. J. (2013). *Handbook of test security*. New York, NY: Routledge.

Students' Perceptions of Their Mathematics Teachers in the Longitudinal Study of American Youth (LSAY): A Factor Analytic Approach

Mohammad Shoraka

Abstract This study investigated the psychometric properties of questionnaire items used to measure students' perceptions of mathematics teachers in the Longitudinal Study of American Youth (LSAY) during middle school. The perceptions of students regarding their math teachers were gathered through 16 questions. The National Science Foundation (NSF) has funded the LSAY and the questionnaire is a collaboration work of the National Centre for Vocational Education Research (NCVER), Department of Education and Training and Wallies Consulting Group. The dataset was randomly split into two samples so that exploratory analyses could be conducted on one-half of the sample and confirmatory analyses could be conducted on the second half. One item, "teacher encourages extra work", was removed from the questionnaire after analyses, due to low loading and ambiguous meaning. Four factors were extracted under different methods of extraction within oblique rotations and were named by the author are: Teachers Characteristics, Teacher Instructional Expectations, Teacher Fairness and Teacher Focus on Outcomes.

Keywords Students' perceptions • Mathematics teachers • Classic Test Theory

1 Introduction

One of the contemporary issues in education of adolescents is their perceptions toward human resources in middle school. The study of students' perceptions has gone beyond the education sector (Balci 2011) with students being viewed as consumers whose parents pay taxes and expect good schooling in return. That is one of the reasons why students' views of teachers or teacher empathy with students have become a topic of interest (Ouazad & Page, 2011). Moreover, finding the dimensional structure of students' perceptions was investigated by conducting a survey and analyzing the resulting data by means of factor analysis.

M. Shoraka (✉)
University of Windsor, Windsor, ON, Canada
e-mail: Shoraka@uwindsor.ca

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer
Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_25

Middle school students' perceptions of parents, peers, families, and teachers are widely viewed as being factors linked to student achievement. For instance, Azmitia, Cooper, and Brown (2009) indicated that the perceptions of students regarding their teachers are not the significant factor of student achievement.

Research suggested that student perceptions of teachers' emphasis on mastery goals predict student self-efficacy. Levpušček and Zupančič (2009) used hierarchical linear modeling analysis and showed that students' perceptions of math teachers' behavior were significant factor on student motivational beliefs as well as their mathematics achievement.

Students by and large spend more of their time with teachers than other academic human resources as they are obligated to attend school regardless of parents' will, in most developed countries. Therefore, there is a legitimate reason why some researchers are concerned about the effect of teacher characteristics on student achievement. Those characteristics could be teacher biases towards student gender and ethnic background (Dee, 2007; Hinnerich, Hoglin, & Johanneson, 2011). Also, teaching experience, performance on state teacher certification exams, certification status and area, competitiveness of a teacher's undergraduate institution, pathway into teaching, and SAT scores on student achievement (Boyd, Lankford, Loeb, Rockoff, & Wyckoff, 2008; Clotfelter, Ladd, & Vigdor, 2007; Neild, Farley-Ripple, & Byrnes, 2009; Olaleye, 2011; Rockoff, 2004).

Many researchers have studied the impact of students' views of teachers' perceptions on students, such as the role of perceived teacher goals on student self-efficacy (Friedel, Cortina, Turner, & Midgley, 2010). Some studies have investigated the effect of student perceptions of receiving support from teacher and classmate on declining attendance (DeWit, Karioja, & Rye, 2010; Nelson-Smith, 2008). Others have studied the effect of teacher expertise on the student's sense of school belonging (Stevens, Hamman, & Olivárez, 2007). In another study, Chen, Thompson, Kromrey, and Chang (2011) investigated the association of teachers' expectations and students' perceptions of teachers' oral feedback in relation to the students.

2 Students' Perceptions of Teachers

Students' perceptions of teachers have been measured according to student-teacher relationships at all levels of education (Cambridge Education, 2012; Hughes, Wu, Kwok, Villarreal, & Johnson, 2012; Koomen, Spilt, & Oort, 2011). For instance, Spilt, Koomen, and Mantzicopoulos (2010) provided children photographs that represented teacher-child interaction and asked children to rate their teachers based upon closeness (e.g., my teacher always listens to me), conflict (e.g., my teacher often gets angry).

By comparison, at the middle school and high school levels, researchers think beyond teacher-student relationships. In his dissertation, Semmel (2007) studied three more factors in addition to student-teacher relationships including: justice,

power and instructional strategies. Sutcliff (2011) also examined two more factors in addition to student-teacher relationship: Justice and Fairness and Instructional Strategies.

3 Purpose of Study

The purpose of this study was to investigate whether or not items of the student survey of mathematics teachers, from grade 7–12 in the Longitudinal Study of American Youth (LSAY) were appropriate for measuring some factors of mathematics teachers' practice. This is a significant study as there have not been any studies evaluating students' perceptions of mathematics teachers in LSAY. Results of this study could be used in future research as one of direct or indirect predictors of students' mathematics achievement.

4 Method

4.1 Instrument

4.1.1 About LSAY

The LSAY has been designed and developed since 1986 to evaluate the progress of student achievement in middle school and high school and its relationship to career. Moreover, the LSAY "was developed to measure student learning about science and technology that might become useful as adults in understanding public policy issues that involve scientific and technological issues" (LSAY, 2011). The LSAY has been funded by the National Science Foundation (NSF).

4.1.2 Data Collection

LSAY researchers collected data in 50 public schools nationally that consisted of two cohorts. The cohort of this study consisted of a national sample of 3116 students starting from the seventh grade over the period of 7 consecutive years. The following was extracted from the user guide of the dataset, explains the sampling data collection:

> The sampling scheme for the base year of the LSAY was a two-stage stratified probability sample. The United States was stratified by four geographic regions and by three levels of urban development (central city, suburban, and non-metropolitan) to produce a total of 12 strata. Stage I [one] involved the selection of schools to participate in the study. Stage II [two] was the random selection of 60 students within each school selected in stage I ... The universe of schools was divided into twelve sampling strata, where the strata were defined

Table 1 Regional sampling of students

Stratum	Target	Initial sample	Response rate	Replacement sample	Total sample
North East					
Urban	120	103	.86	18	121
Suburban	300	213	.71	134	347
Rural	120	112	.93	38	150
North Central					
Urban	240	208	.87	45	253
Suburban	360	335	.93	21	356
Rural	299	288	.96	53	341
South					
Urban	300	266	.89	35	301
Suburban	360	332	.92	37	369
Rural	349	315	.90	34	349
West					
Urban	120	101	.84	20	121
Suburban	300	260	.87	35	295
Rural	119	107	.90	5	112

by the cross classification of region of the country (Northeast, Midwest, South, and West) and community type (urban, suburban, and rural) . . . students were selected randomly from the lists and asked to participate until the target response size was achieved (ICPSR, 2011, p. 5).

4.1.3 Participants

Grade 7 students were divided into four regions: North East, North Central, South and West (see Table 1).

4.2 Data Analysis

The perceptions of students regarding their math teachers were gathered through 16 questions administered in the spring of the school year (see Table 2).

Each question has three options in the following order: true, false and not sure. In order to use the scale properly the questionnaire was re-coded so the choice of 'not sure' was put in the middle. In other words, a higher score, such as three, means the teacher had good characteristics, and as it goes lower, those characteristics decreased in quality. However, three items had a negative meaning and needed to be reversed coded.

The following choices caused missing values: multiple punches, blank response or lack of a course. The total number of missing cases was small, between 0.7 and

Table 2 Items of each factor

Name of the factor	Item	Item	Item	Item
Teacher characteristics	Likes me	Enjoy teaching	Very good teacher	Gives help
Teacher instructional expectations	Expects completed homework	Expects hard work	Expects best from me	Thinks I should do well
Teacher focus on outcomes	Encourages math/science career	Expects go to college	Encourages me in math	Talks about jobs
Teacher fairness	Treats boys and girls differently	Pays more attention to boys	Makes me feel dumb	

1.7 % of the data so listwise deletion was applied for the analysis (Graham & Hofer, 2000). Moreover, students who did not participate were not counted as the missing cases.

The dataset was randomly split into two samples so that exploratory analyses could be conducted on one-half of the sample and confirmatory analyses could be conducted on the second half. One file is used for Exploratory Factor Analysis (EFA) with 1472 cases and another for Confirmatory Factor Analysis (CFA) with 1368 cases. In order to get a good sense of the number of factors, it is good to use a different number of factors with multiple methods of extraction. For EFA, I expected the factors to be correlated so I rotated the initial solution using an oblique rotation procedure, Promax. The Promax solution was easiest to interpret among the several oblique rotation procedures I tried.

4.2.1 EFA Results

EFA was deployed through Mplus and items were treated as categorical variables. Four factors were extracted under different methods of extraction within oblique rotations. Factors were named by the author: Teacher Characteristics, which included four items; Teacher Instructional Expectations, which included four items; Teacher Focus On Outcomes (Career, Post-Secondary Education, Math Classes), which included four items; and Teacher Fairness (TF), which included three items. Table 2 shows items of each factor.

4.2.2 CFA Results

I used the results of EFA on the second half of dataset and applied CFA to investigate the validity of the study. In addition to the Chi-Square, the model was examined through two measures suggested by Hu and Bentler (1999): Comparative Fit Index (CFI) greater or equal to .95, Root Mean Squared Error of Approximation (RMSEA)

less than or equal to .06. Chi-Square was significant at 539 with 84 degrees of freedom, but CFI was equal to .938 and RMSEA was at .061, suggesting that the model is close to the acceptable measures.

5 Results and Discussions

Other than one item, 'teacher encourages extra work', which had low loadings, other items had loadings higher than 0.4 (see Table 3). The item with a low loading has a problem with the word 'extra work', which could be interpreted in several ways. For instance, 'extra work' might mean 'beyond the current assignment' and refers to the future, which relates to the Teacher Focus on Outcomes, or it could mean 'extra in addition to current task', which relates to the Teacher Instructional Expectations, or it could mean just in general the teacher encourages extra work, which relates to the Teacher Characteristics. Moreover, the phrase 'extra work' could be positive to some students and negative to others. It could mean positive to students if extra work meant students worked well before and the teacher gives them extra to advance their knowledge. However, extra work could interpret negatively if extra work is unrewarded. Thus, item 'encourages extra work' was excluded from the analysis. Cognitive interview and logical analysis theory should be used in addition to statistical analysis for this item.

Another noticeable item is 'teacher makes me feel dumb' because it has a higher loading for the Teacher Characteristics factor than Teacher Fairness; however, this item, along with the two other items 'teacher treats boys and girls differently' and 'teacher pays more attention to boys', have negative connotations. Moreover, when a student reads 'teacher makes me feel dumb' it implies that this is not fair. The assumption is that the student does not think he or she is dumb unless somebody makes him or her feeling that way. Thus, it is more appropriate to consider this as part of the Teacher Fairness factor.

Table 4 shows the descriptive statistics for the 16 items. Table 5 shows the descriptive of the four factors and Table 6 shows the correlation among the four factors. The kurtosis of Teacher Instructional Expectations was noticeable at 3.70; however, according to Kline (2005), that number could be even higher, up to 10, and still be acceptable.

Among these four factors, two factors were common with the two other studies; those were Instructional Strategies and Justice in Semmel (2007), and Instructional Strategies and Justice and Fairness in Sutcliff (2011). Cronbach alphas of Teacher Instructional Expectations and Teacher Fairness for this study were .74 and .62, respectively, in line with two studies mentioned above for the instructional factor, (.72–.83) and for the fairness factor, (.51–.63). See Table 7 for the reliability of the four factors.

Table 3 Promax rotated loadings (n = 1472) weighted least squares extraction, promax rotation, pattern coefficients

Items factors	Teacher instructional expectations	Teacher characteristics	Teacher fairness	Teacher focus on outcomes
Enjoys teaching	.16	.57	−.02	.08
Expects best from me	.85	.07	−.03	−.06
Encourages extra work	.296	.232	−.125	.219
Expects hard work	.96	.03	−.08	−.12
Expects completed homework	.79	−.08	.17	−.04
Very good teacher	.02	.90	.03	−.09
Talks about jobs	−.22	.06	−.15	.74
Expects us to go to college	.18	−.01	.04	.47
Encourages math	.23	.01	.03	.63
Encourages math or science career	−.15	−.10	.10	.88
Thinks I should do well	.40	.30	.13	.11
Treats boys and girls differently	−.09	.16	.77	.01
Makes me feel dumb	−.10	.46	.38	−.12
Pays more attention to boys than girls	.07	−.12	.90	.03
Gives extra help	.12	.64	−.04	.07
Really likes me	−.08	.77	.02	.08

6 Conclusions

LSAY is one of the most valuable resources for many analysts. The numbers of dissertations that have come out of the LSAY survey have reached 31. The fact that their participants could be tracked over 25 years is remarkable. The results of this study have proposed some questions to researchers and have provided a cautious recommendation as well. One section of the LSAY survey relates to students who were in grade 7 and asked the same questions on their perceptions of mathematics teachers until they reached grade 12.

The outcome of EFA and CFA indicate that the questionnaire measures four factors and those factors are Teacher Characteristics, Teacher Instructional

Table 4 Descriptive statistics of the 16 items (n = 1472)

Item	M	SD	Skewness	Kurtosis
Enjoy teaching	2.66	0.58	−1.48	1.17
Pays more attention to boys	2.70	0.63	−1.88	2.08
Treats boys and girls differently	2.50	0.79	−1.15	−0.43
Gives extra help	2.43	0.86	−0.96	−0.95
Likes me	2.32	0.76	−0.61	−1.05
Very good teacher	2.57	0.73	−1.36	0.22
Makes me feel dumb	2.61	0.74	−1.54	0.58
Expects best from me	2.73	0.60	−2.09	2.96
Expects hard work	2.77	0.56	−2.34	4.21
Expects completed homework	2.83	0.52	−2.94	7.18
Thinks I should do well	2.70	0.62	−1.90	2.24
Talks about jobs	1.32	0.69	1.83	1.61
Expects go to college	2.04	0.74	−0.06	−1.19
Encourages extra work	2.13	0.91	−0.26	−1.75
Encourages math/science career	1.53	0.76	1.03	−0.53
Encourages me in math	1.97	0.90	0.06	−1.77

Note: Item response scale could range from 1 to 3

Expectations, Teacher Fairness, Teacher Characteristics and Teacher Focus on Outcomes (career, post-secondary education and math in general). The reliabilities of these factors were moderate and close to other studies (Semmel, 2007; Sutcliff, 2011). These four factors could be used by future researchers as the predictors for any analysis in LSAY.

7 Limitations of Study and Future Research

There are some studies about students' perceptions of their teachers in general, but not in math teachers in particular. Nonetheless, using this survey in a longitudinal study could be misleading for three reasons: the first reason is that the questions about math instructional approaches are very general. The second relates to the amount of missing data particularly in grade 12 in two particular items. Those items are as follows: math teacher 'gives extra help' and 'really likes me'. The third reason pertains to another item 'encourages extra work' that seems problematic as it loads in three factors. If that item (math teacher 'encourages extra work') is removed, the results could be cautiously reliable. Further research on other subjects that are available in the database such as science is recommended. The growth model analysis could also shed more light on the student perceptions of math teacher.

Acknowledgment I would like to thank Professor Robert Dedrick for reading, editing and commenting on this study.

Table 5 Descriptive statistics for the four factors (n = 1472)

Factor	# of items	M	SD	Skewness	Kurtosis	Item-to-total correlation	Cronbach's alpha
Teacher characteristics	4	2.45	0.57	−0.95	−0.04	.72−.77	.74
Teacher fairness	3	2.62	0.53	−1.31	0.07	.59−.66	.62
Teacher instructional expectations	4	2.74	0.44	−1.99	3.70	.71−.76	.74
Teacher focus on outcomes	4	1.67	0.52	0.67	−0.12	.59−.66	.63

Table 6 Factor correlation matrix for grade 7 (n = 1472) weighted least squares extraction, promax rotation

Factor	Teacher instructional expectations	Teacher characteristics	Teacher fairness	Teacher focus on outcomes
Teacher instructional expectations	1.000			
Teacher characteristics	.61	1.000		
Teacher fairness	.30	.45	1.000	
Teacher focus on outcomes	.33	.43	−.24	1.000

Table 7 Reliability scores for each subscale with or without a low loading item (n = 1472)

Subscale	Cronbach alpha
Teacher Fairness (TF)	.62
Teacher Characteristics (TC)	.74
Teacher Focus On Outcomes (TFOO)	.63
Teacher Instructional Expectations (TIE)	.74

References

Azmitia, M., Cooper, C. R., & Brown, J. R. (2009). Support and guidance from families, friends, and teachers in Latino early adolescents' math pathways. *The Journal of Early Adolescence, 29*(1), 142–169.

Balci, A. (2011). Perceptions of secondary level students about the United Nations (UN) and the permanent member states. *European Journal of Educational Studies, 3*(3), 429–452.

Boyd, D., Lankford, H., Loeb, S., Rockoff, J., & Wyckoff, J. (2008). The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools. *Journal of Policy Analysis and Management, 27*(4), 793–818.

Cambridge Education. (2012). *Tripod survey assessment*. Retrieved from http://www.camb-ed-us.com/Home.aspx.

Chen, Y., Thompson, M. S., Kromrey, J. D., & Chang, G. H. (2011). Relations of student perceptions of teacher oral feedback with teacher expectancies and student self-concept. *The Journal of Experimental Education, 79*(4), 452–477.

Clotfelter, T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review, 26*(6), 673–682.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources, 42*(3), 528–554.

DeWit, D. J., Karioja, K., & Rye, B. J. (2010). Student perceptions of diminished teacher and classmate support following the transition to high school: Are they related to declining attendance? *School Effectiveness and School Improvement, 21*(4), 451–472.

Friedel, J. M., Cortina, K. S., Turner, J. C., & Midgley, C. (2010). Changes in efficacy beliefs in mathematics across the transition to middles school: Examining the effects of perceived teacher and parent goal emphases. *Journal of Educational Psychology, 102*(1), 102–114.

Graham, J. W., & Hofer, S. M. (2000). Multiple imputation in multivariate research. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multiple group data: Practical issues, applied approaches, and specific examples*. Hillsdale, NJ: Erlbaum.

Hinnerich, B. T., Hoglin, E., & Johanneson, M. (2011). *Ethnic discrimination in high school grading: Evidence from a field experiment*. Scandinavian Working Papers in Economics. Working Paper Series in Economics and Finance No. 733. Retrieved from http://swopec.hhs.se/hastef/papers/hastef0733.pdf.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Hughes, J. N., Wu, J., Kwok, O., Villarreal, V., & Johnson, A. Y. (2012). Indirect effects of child reports of teacher–student relationship on achievement. *Journal of Educational Psychology, 104*(2), 350–365.

ICPSR. (2011). *Longitudinal study of American youth, 1987–1994, 2007–2011*. Retrieved from http://www.icpsr.umich.edu/icpsrweb/NACJD/studies/30263.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: The Guilford Press.

Koomen, H. M. Y., Spilt, J. L., & Oort, F. J. (2011). The influence of affective teacher–student relationships on students' school engagement and achievement: A meta-analytic approach. *Review of Educational Research, 81*(4), 493–529.

Levpušček, M. P., & Zupančič, M. (2009). Math achievement in early adolescence: The role of parental involvement, teachers' behavior, and students' motivational beliefs about math. *The Journal of Early Adolescence, 29*(4), 541–570.

LSAY. (2011). *About LSAY: History and mission*. Retrieved from http://lsay.org/about_history.html.

Neild, R. C., Farley-Ripple, E. N., & Byrnes, V. (2009). The effect of teacher certification on middle grades achievement in an urban district. *Educational Policy, 23*(5), 732–760.

Nelson-Smith, F. N. (2008). *Learning styles and students' perception of teachers' attitudes and its relation to truancy among African American students in secondary education*. Unpublished doctoral dissertation, Louisiana State University, Baton Rouge, LA.

Olaleye, F. O. (2011). Teachers characteristics as predictor of academic performance of students in secondary schools in Osun State. *European Journal of Educational Studies, 3*(3), 505–511.

Ouazad, A., & Page, L. (2011). *Students' perceptions of teacher biases: Experimental economics in schools*. Social Science Research Network. INSEAD Working Paper No. 2011/88/EPS. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1782675##.

Rockoff, J. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review, 94*(2), 247–252.

Semmel, M. J. (2007). *The association between high school students' perception of teacher influenced variables and students' perception of engagement*. Doctoral dissertation, University of Connecticut.

Spilt, J. L., Koomen, H. M. Y., & Mantzicopoulos, P. Y. (2010). Young children's perceptions of teacher–child relationships: An evaluation of two instruments and the role of child gender in kindergarten. *Journal of Applied Developmental Psychology, 31*, 428–438.

Stevens, T., Hamman, D., & Olivárez, A., Jr. (2007). Hispanic students' perception of white teachers' mastery goal orientation influences sense of school belonging. *Journal of Latinos and Education, 6*(1), 55–70.

Sutcliff, C. P. (2011). *Secondary students' perceptions of teacher qualities.* Doctoral dissertation, Georgia Southern University.

Influential Factors of China's Elementary School Teachers' Job Satisfaction

Hong-Hua Mu, Mi Wang, Hong-Yun Liu, and Yong-Mei Hu

Abstract Using nationwide representative data from the Chinese National Innovation Center for Assessment (CNICA) of Basic Education Quality in 2014, this study investigated the influential factors of China's elementary school teachers' job satisfaction (TJS), and identified factors that contribute to it. The following findings were obtained from a series of hierarchical linear models. The TJS was the highest in the urban areas, and the lowest in rural areas, with town areas in between. There was a significant gap in TJS among schools. In addition to the professional title, other demographic variables such as gender, years of teaching, and educational background, had statistically significant but practically small impacts on TJS in elementary schools. Teachers' daily workload, monthly income, and part-time job situation could significantly predict TJS, and had a greater influence than the demographic variables. Teachers' self-directed professional development, occupational preferences, and job engagement had significantly effects on the prediction of TJS, and the effects were greater than those of the objective factors. Besides the institutional culture factors which contributed to negative effects such as encouragement of teaching innovation and teaching supervision, principal's instructional leadership, teachers' professional development support, and democratic decision-making played significantly positive roles and the influence were greater than factors from the teachers. The factors affecting TJS in elementary

H.-H. Mu, Ph.D. (✉)
College of Education Administration & Collaborative Innovation Center of Assessment towards Basic Education Quality, Beijing 100875, China
e-mail: 201431010066@mail.bnu.edu.cn

M. Wang
University of Maryland College Park, College Park, MD 20904, USA

H.-Y. Liu
School of Psychology & Collaborative Innovation Center of Assessment towards Basic Education Quality, Beijing 100875, China

Y.-M. Hu
College of Education Administration & Collaborative Innovation Center of Assessment towards Basic Education Quality, Beijing Normal University, Beijing 100875, China

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer
Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_26

schools from more important to less important were: professional development support, principal's instructional leadership, occupational preferences, and job engagement.

Keywords Professional development • Teacher job satisfaction • Principal's instructional leadership • Hierarchical linear models

1 Introduction

All nations around the world hold education high in esteem to achieve economic growth. From the past world education trends, especially the last few decades, we have no difficulty in drawing the conclusion that teachers are the critical roles to enhance education quality, therefore, how to improve teacher's quality is the key to break through in each country.[1] Many studies have demonstrated that teachers are not only the core pillars of the school development, but more importantly, the executives of the education quality promotions (Miao 2006). Some studies manifested that teachers' degrees of job satisfaction have high impacts on job performance, career identity, affective commitment, job attitude, and organizational behavior, as well as their senses of the collective efficacy and self-efficacy, and teaching efficacy (Arifin 2015). In addition, some researchers stated that "Identification of factors influence teacher job satisfaction would have positive impacts on teachers' quality and professional identity, as well as student's academic achievement, and education satisfaction rate" (Bogler 2002; Klassen & Tze 2014; Wang, Hall & Rahimi 2015). Moreover, there is a substantial connection between teacher job satisfaction and their intention to quit the job (Liu & Meyer, 2005). For teachers, the lower the job satisfaction is, the higher intention to quit or change to another career. Under the situation of China's urbanization acceleration, there is significant loss of prominent teachers in villages, towns, and other rural areas (Li 2013).

In this era of pursuing well-balanced educational development and high education quality, the investigation of key factors affecting TJS in elementary schools can contribute to inspiring teachers' work enthusiasm, stabilizing teaching faculty (especially in rural elementary schools), and achieving a sustainable development of education. It is therefore very important to identify factors affecting TJS, which can advise education authorities and school administrators to make appropriate decisions to improve educational quality.

[1] From 2008 to 2014, OECD countries education policy classification statistics show that the policy of supporting school improvement accounted for 24 %, and these policies focused on promoting teachers' professional development, curriculum reform. For details, the reader is referred to the OECD Education Briefing: Education Policy Outlook 2015 and Education at a Glance. http://www.oecd.org/edu/eag.htm www.oecd.org/edu/eag-interim-report.htm-2015-1-20.

2 Literature Review

The research of job satisfaction began in the early eighteenth century. Taylor claimed that "high pay can improve job satisfaction," which was followed by a wave of "job satisfaction" researches. On job satisfaction, many studies focused on the structure and affecting factors. In this section, we review the concepts and influential factors of TJS.

2.1 Concepts of Teacher Job Satisfaction

In the field of management, sociology, and psychology, job satisfaction is an important indicator of how employees feel about their jobs. Hoppock (1935) claimed that job satisfaction includes mental, physical, and environmental satisfaction of employee. Herzberg (1968) described job satisfaction as job motivation. Locke (1976) insisted that job satisfaction is a pleasurable or positive emotional state coming from the job. According to these studies, TJS can be referred to as the overall subjective and emotional feelings and opinions that teachers have towards their occupations and working conditions. In some other studies, TJS was described as the fulfillment he/she drives from daily jobs (Korb & Akintunde 2013), or as a teacher's affective relation to his or her teaching role and is a function of the perceived relationship between what one wants from teaching and what one perceives it is offered to a teacher (Zembylas & Papanastasiou 2004). All in all, according to Luthans, Zhu, and Avolio (2006), TJS can be comprehended as a kind of teachers' responses to their working conditions and environments.

2.2 Influential Factors of Teacher Job Satisfaction

For the structure of job satisfaction, despite different opinions of researchers, the contents are similar. The representative studies led by Herzberg (1968) and Friedlander (1964) focused on two-factor theory and three-factor theory, respectively. Herzberg believed that job satisfaction is mainly consisted of motivators and hygiene factors, while Friedlander believed it contains three factors, social and technical environment, recognition factors, and self-realization factors. Environmental factors include superior-subordinate communication, working conditions, interpersonal relationships, etc.; recognition factors consist of work challenging, income, responsibilities, promotion, etc.; self-realization factors include those where individual abilities could get on plays. Therefore, the structure of job satisfaction can be summed up as working conditions and environmental factors (including superior-subordinate communication, interpersonal relationships, working conditions and so on), recognition factors (responsibility, promotion, etc.), and self-realization factors.

According to the structure of job satisfaction and past research, there are three kinds of influential factors of TJS:

(a) Teacher Variables

Teachers' demographic variables consist of gender, age, years of teaching (teaching experience, length of service), level of education, and job title (Chen & Sun 1994; Hu 2007). Chen and Sun (1994) found that TJS differed between genders, and female teachers had a higher satisfaction than their male colleagues. On the contrary, Feng (1996) found that in Beijing middle schools, male teachers had a higher TJS, whereas Hu (2007) suggested that gender had no significant influences on TJS.

Crossman and Harris (2006) suggested no significant difference in UK secondary school TJS between ages, genders, and lengths of service. Mukhopadhyaya and Kabirak (2014) suggested no significant difference between male and female teachers. Iqbal and Akhtar (2012) observed that female teachers were more satisfied with work, while age and work experience did not play a role on job satisfaction. Wu (1996) reported a significant effect of age and teaching level.

Objective variables (work itself) consist of work condition, job stress, job involvement, income, and advancement, etc. Wu (1996) found that teachers in the 26–30 years' age group reported the lowest level of satisfaction on advancement. Li (2013) found that daily work load had no significant effect on new teachers' job satisfaction. Korb and Akintunde (2013) found job satisfaction was not related to salary. Subjective variables consist of self-directed professional development, teacher efficacy, occupational preferences, personality, etc. Li (2013) found that teacher enthusiasm, teaching ability, and demand for advancement all had a significant impact on beginning teachers' job satisfaction. Skaalvik and Skaalvik (2010) found that teacher self-efficacy was related to TJS.

(b) School Variables

The variables include school location, school type, school administration, principal's leadership, interpersonal relationship, supervision, school environment, school culture, and democratic decision-making, etc. Iqbal and Akhtar (2012) found no significant difference in TJS between urban and rural school teachers, whereas Mukhopadhyaya and Kabirak (2014) observed a significant difference. Wu (1996) reported a significant effect between school types, and a rather high level of satisfaction with supervision. Shen, Leslie, Spybrook, and Ma (2012) recommended that principal's leadership had significant effect on TJS. Li (2013) found school location and institutional environment factors had a significant effect on the beginning teachers' job satisfaction. Skaalvik and Skaalvik (2010) found that teacher self-efficacy was related to school context variables. Aldridge and Fraser (2015) found that TJS was related to school climate.

(c) Student Variables

The variables include positive student behavior, teacher-student relationship, academic achievements, etc. Veldman, van Tartwijk, Brekelmans, and Wubbels (2013) indicated that TJS appeared positively related to the self-reported quality of the teacher-student relationships. Caprara, Barbaranelli, Steca, and Malone (2006) indicated that students' previous academic achievement did not contribute to TJS. Shen et al. (2012) recommended that positive student behavior was associated with TJS.

In a word, the studies on TJS in the literature are inconclusive. Most studies in China focus more on teachers in less-developed western regions or cities in developed regions, whereas few attentions are given to teachers in rural areas or urban-rural contrast. To increase the comparability between China and other countries and to understand the status of TJS in elementary schools and explore the key influential factors of teacher job satisfaction, we implemented a nationwide random survey to obtain a representative sample. Following to practices in the previous studies, we classified the factors into two categories: individual factors and organizational factors. Individual factors have received much research attention than organizational factors (Judge et al. 2002). In this study we hypothesized that school institutional culture (decision-making, principal leadership, etc.) plays an important role in TJS. Based on the authoritative survey from the CNICA of Basic Education Quality towards Basic Education Quality's regional projects and the use of HLM, this study sought to identify influential factors of China's elementary school teachers' job satisfaction. Specifically, we aimed to answer the following three questions: (a) To what extent are the differences in TJS amongst elementary schools in China? (b) To what extent are the differences in TJS amongst different areas, such as urban, township, rural areas? (c) After controlling for school location variables and teachers' demographic variables, to what extent do teachers level factors and school level factors account for the variations in TJS?

3 Research Methodology

3.1 Data Sources and Sample Composition

The CNICA towards Basic Education Quality's regional projects adopted probability proportional to size to collect data. Respondents were 13,406 teachers from 2019 elementary schools (542 rural schools, 569 township schools, 825 urban schools, 83 schools' information missing) in 163 districts and counties from Shijiazhuang, Xilin Gol League, Zhejiang Province, Zhengzhou, Luoyang, Zhuzhou, and Shenzhen. Among those schools, male teachers accounted for 25.2 %, female teachers 74.7 %, and 0.1 % without gender information; teachers without a bachelor's degree took up to 26.1 %, and teachers with a bachelor's degree accounted for 72.1 %, while 1.7 % had a master's or higher degree, and 0.1 % without information;15.3 % had less than

5 years of teaching experience, 15.4 % had 5–10 years, and 37 % had 11–20 years, and 32.3 % had 20 years or more. For the job title, 40.3 % were junior, 56.8 % were senior or higher, and 2.9 % without information.

3.2 Variables and Measurement

3.2.1 Teacher Job Satisfaction Scale

The Teachers Job Satisfaction Scale was a self-reported rating scale revised from Hu's (2007) TJS questionnaire, and the scale consisted of 16 questions on 5 dimensions (school leadership and management, professional development environment, effort-reward, interpersonal relationships, and self-actualization) with five points (very unsatisfied, relatively unsatisfied, not sure, quite satisfied, very satisfied). The average score for each dimension was used as a metric for each dimension, and the average score across the five dimensions reflected a teacher's overall evaluation, with a higher average score indicating a higher level of satisfaction.

Among these five dimensions, leadership and management satisfaction and development of the environment satisfaction had the highest correlation ($r = 0.694$), effort-reward reasonability satisfaction and interpersonal relationships satisfaction have the lowest correlation coefficient ($r = 0.284$), and the other correlations were about 0.5. Validity and reliability of the scale were relatively good, with $\alpha = 0.905$ and factor loadings by confirmatory factor analysis for the five dimensions of 0.771–0.894, 0.655–0.840, 0.700–0.800, 0.531–0.780, and 0.619–0.762, respectively.

3.2.2 Teachers' Individual Level Explanatory Variables

3.2.2.1 Objective Variables from Teachers

Objective variables were obtained from the teacher questionnaires as follows. Daily workload was coded as: 5–7 h = 1, 8–9 h = 2, 10–11 h = 3, 12 h and above = 4, and missing = 99. The percentages of these categories were 4.7 %, 51.1 %, 35.3 %, 8.5 %, and 0.3 %, respectively. Part-time job situation was coded as: with part-time job = 1, without part-time job = 2, Missing = 99. The percentages of these three categories were 56.6 %, 43.1 %, and 0.3 %, respectively. Monthly income was coded as 3000 yuan = 1, 3000–4000 yuan = 2, 4000–5000 yuan = 3, and 5000 yuan above = 4. The percentages of these categories were 34.9 %, 30.1 %, 20.5 %, and 14.5 %, respectively.

3.2.2.2 Subjective Variables from Teachers

Occupation preferences were originated from the teacher questionnaires, and the coding was: do not like = 0, generally like = 1, very much = 2. The percentages of these three categories were 13.2 %, 56.9 %, and 29.8 %, respectively. Teacher job involvement was originated from the Engagement Scale in the teachers' questionnaires. The scale was revised mainly based on the Utrecht Work Engagement Scale developed by Schaufeli and Bakker (2003), and consisted of nine items on three dimensions of teaching vigor, teaching recognition, and teaching reflection. The scale had five points: not at all, not really, undecided, somewhat, very much. The study used the average score of the three dimensions as an overall evaluation of teachers' job commitment, and a higher score indicated a higher job involvement. Correlation coefficients between work vitality and work identity and between work vitality and work attentiveness were 0.406 and 0.573, respectively; while correlation between work identity and work attentiveness was 0.393. The validity and reliability of the scale were quite favorable with $\alpha = 0.761$ and factor loadings of three dimensions of 0.518–0.705, 0.601–0.759, and 0.621–0.692, respectively.

Teachers' self-directed professional development was originated from the professional development scale of the teachers' questionnaire. The scale had seven questions on three dimensions: professional guide and innovation, colleague communication and mutual assistance, and teaching reflection, with five points: never, occasionally, sometimes, often, always. The study used the average score of three dimensions as an overall evaluation of teachers' self-directed professional development, with a higher average score indicating a better professional development. Correlation coefficients of professional guidance and innovation to colleagues' communication and mutual assistance, and teaching reflection were 0.545 and 0.416, respectively, while the correlation between the latter two was 0.597. The validity and reliability of the scale were quite favorable, with $\alpha = 0.810$ and factor loadings of the three dimensions of 0.471–0.740, 0.753–0.775, and 0.661–0.886, respectively.

3.2.2.3 System Environment Variables at the School Level

The evaluation of organizational environment of school was based on 25 questions of 5 dimensions: democratic decision-making, innovative teaching encouragement, teachers' professional development support, teaching supervision, and principal instructional leadership. The scale had five points: not at all, not really, undecided, somewhat, very much. A higher average score indicated a better environment. The validity and reliability of the scale were relatively favorable. There were 4 questions in democratic decision-making, with $\alpha = 0.825$ and factor loadings of 0.753–0.859. The dimension of teaching innovation encouragement had 4 questions, with $\alpha = 0.878$ and factor loadings of 0.747–0.888. The dimension of teaching supervision had 3 questions, with $\alpha = 0.613$ and factor loadings of 0.464–0.668. The dimension of professional development support had five questions, with $\alpha = 0.888$

Table 1 Descriptive statistics of regional teachers' dimensions and overall job satisfaction

Location	Satisfaction	*n*	*M*	*SD*
Urban	Leadership and management	4107	4.039	0.786
	Environment for development	4107	4.025	0.802
	Reasonableness of effort—reward	4107	2.555	1.011
	Self-fulfilling	4107	3.673	0.794
	Interpersonal relationships	4106	4.236	0.579
	Teacher job satisfaction	4107	3.706	0.608
Township	Leadership and management	4422	4.044	0.773
	Environment for development	4421	3.97	0.791
	Reasonableness of effort-reward	4421	2.402	1.016
	Self-fulfilling	4419	3.636	0.802
	Interpersonal relationships	4419	4.236	0.575
	Teacher job satisfaction	4422	3.657	0.599
Rural areas	Leadership and management	4727	3.998	0.754
	Environment for development	4725	3.873	0.775
	Reasonableness of effort-reward	4727	2.43	0.939
	Self-fulfilling	4727	3.603	0.763
	Interpersonal relationships	4727	4.192	0.548
	Teacher job satisfaction	4727	3.619	0.562

and factor loadings of 0.712–0.863. The dimension of principal instructional leadership had 5 questions, with $\alpha = 0.886$ and factor loading of 0.770–0.902. The correlations among the dimensions are shown in Table 1.

3.3 Data Processing

3.3.1 Analysis Tools

The computer programs for data analyses in this study were SPSS 17.0 and Mplus 7.0.

3.3.2 Analysis Methods

Since the data has a nested structure (teachers were nested in schools), this study used HLM to explore the influences of the independent variables at different levels on TJS.

3.3.3 Specific Analysis Steps

3.3.3.1 Elementary Teachers' Job Satisfaction Among Different Schools (Model_0)

*Model*_0:

$$\text{Teacher-level}: \quad \text{TJS}_{ij} = \beta_{0j} + r_{ij}, r_{ij} \sim N\left(0, \delta^2\right) \tag{1}$$

$$\text{School-level}: \quad \beta_{0j} = \gamma_{00} + \mu_{0j}, \mu_{0j} \sim N\left(0, \tau_{00}\right) \tag{2}$$

Firstly, we established a baseline model without any forecast variable of elementary school teachers' job satisfaction, to decompose the total variations of TJS into two different levels: within-school and between-school, to study whether there were within-school variations on elementary school TJS. The specific model is as follows:

*Model*_0:

$$\text{Teacher-level}: \quad \text{TJS}_{ij} = \beta_{0j} + r_{ij}, r_{ij} \sim N\left(0, \delta^2\right) \tag{1}$$

$$\text{School-level}: \quad \beta_{0j} = \gamma_{00} + \mu_{0j}, \mu_{0j} \sim N\left(0, \tau_{00}\right) \tag{2}$$

Among them, TJS_{ij} is the score on TJS for teacher j in the school i, β_{0j} is the average score of teacher job satisfaction, γ_{ij} is the teacher-level random effect, δ^2 is the variability within schools; γ_{00} is the grand mean (or intercept), μ_{0j} the school-level random effect, τ_{00} is the variability across schools.

3.3.3.2 The Effects the Demographic Variables of Teachers and School Location (Model_1)

In order to explore the effects of the demographic variables of teachers' and school location on TJS, the following variables were added to Model_0: gender, years of teaching, job title, education background, and school location, respectively, to form Model _1 as follows.

*Model*_1:

$$\begin{aligned}\text{Teacher-level}: \quad \text{TJS}_{ij} = {} & \beta_{0j} + \beta_{1j}(gender) + \beta_{2j}\,(education\ \ background) \\ & + \beta_{3j}\,(years\ \ of\ \ teaching) + \beta_{4j}\,(job\ \ title) \\ & + r_{ij}, r_{ij} \sim N\left(0, \delta^2\right)\end{aligned} \tag{3}$$

$$\text{School-level}: \quad \beta_{0j} = \gamma_{00} + \gamma_{01}(rural) + \gamma_{02}(town) + \mu_{0j}, \mu_{0j} \sim N\left(0, \tau_{00}\right) \tag{4}$$

β_{1j}–β_{4j} are the coefficients (effects) of the covariates on TJS; other parameters have been defined in the baseline model; γ_{01} and γ_{02} are the coefficients (effects) of the covariates on the intercept.

3.3.3.3 Effects of Teacher's Objective Variables (Model_2)

Model_2 was developed by adding objective factors from teachers (daily workload, monthly income, and homeroom teacher situation) to Model_1. Model_2 was mainly used to examine the effects of objective variables on elementary school TJS, which is specified as follows.

Model_2:

$$\begin{aligned}\text{Teacher-level}: \quad \text{TJS}_{ij} = {} & \beta_{0j} + \beta_{1j}(gender) + \beta_{2j}\,(education\ \ background) \\ & + \beta_{3j}\,(years\ \ of\ \ teaching) + \beta_{4j}\,(job\ \ title) \\ & + \beta_{5j}\,(daily\ \ workload) + \beta_{6j}\,(monthly\ \ income) \\ & + \beta_{7j}\,(part-time) + r_{ij}, r_{ij} \sim N\left(0, \delta^2\right) \end{aligned} \tag{5}$$

$$\text{School-level}: \quad \beta_{0j} = \gamma_{00} + \gamma_{01}(rural) + \gamma_{02}(town) + \mu_{0j}, \mu_{0j} \sim N\,(0, \tau_{00}) \tag{6}$$

β_{5j}–β_{7j} are the coefficients (effects) of the covariates on TJS, other parameters have been defined in the baseline model.

3.3.3.4 Effects of Teachers' Subjective Variables (Model_3)

Model_3 was created by adding teachers' subjective variables (occupational preferences, job involvement, and self-professional development) to Model_2. Model_3 was mainly used to examine effects of teachers' subjective variables on TJS, as shown as follows.

Model_3:

$$\begin{aligned}\text{Teacher-level}: \quad \text{TJS}_{ij} = {} & \beta_{0j} + \beta_{1j}(gender) + \beta_{2j}\,(education\ \ background) \\ & + \beta_{3j}\,(years\ \ of\ \ teaching) + \beta_{4j}\,(job\ \ title) \\ & + \beta_{5j}\,(daily\ \ workload) + \beta_{6j}\,(monthly\ \ income) \\ & + \beta_{7j}\,(part-time) + \beta_{8j}\,(occupation\ \ preference) \\ & + \beta_{9j}\,(job\ \ involvement) + \beta_{10j}\,(self-directed\ \ \text{PD}) \\ & + r_{ij}, r_{ij} \sim N\left(0, \delta^2\right) \end{aligned} \tag{7}$$

$$\text{School-level}: \quad \beta_{0j} = \gamma_{00} + \gamma_{01}(rural) + \gamma_{02}(town) + \mu_{0j}, \mu_{0j} \sim N\,(0, \tau_{00}) \tag{8}$$

β_{8j}–$\beta_{(10)j}$ are the coefficients (effects) of the covariates on TJS, other parameters have been defined in the baseline model.

3.3.3.5 Effects on School Organizational Environment (Model_4)

Model_4 was established by adding the explanatory variables at the school level (teaching innovation encouragement, professional development support, teaching supervision, principal instructional leadership, and participation in decision making) to Model_3. This model was mainly used to examine the effects of school environment on elementary school TJS, as shown below.

*Model*_4:

$$\begin{aligned}\text{Teacher-level}: \quad \text{TJS}_{ij} = {} & \beta_{0j} + \beta_{1j}(gender) + \beta_{2j}\,(education\ \ background) \\ & + \beta_{3j}\,(years\ \ of\ \ teaching) + \beta_{4j}\,(job\ \ title) \\ & + \beta_{5j}\,(daily\ \ workload) + \beta_{6j}\,(monthly\ \ income) \\ & + \beta_{7j}\,(part-time) + \beta_{8j}\,(occupation\ \ preference) \\ & + \beta_{9j}\,(job\ \ involvement) \\ & + \beta_{10j}\,(self-direct\ \ professional\ \ development) \\ & + r_{ij}, r_{ij} \sim N\left(0, \delta^2\right) \end{aligned} \tag{9}$$

$$\begin{aligned}\text{School-level}: \quad \beta_{0j} = {} & \gamma_{00} + \gamma_{01}(rural) + \gamma_{02}(town) \\ & + \gamma_{03}\,(democratic\ \ decision-making) \\ & + \gamma_{04}\,(professional\ \ development\ \ support) \\ & + \gamma_{01}\,(teaching\ \ supervision) \\ & + \gamma_{01}\,(principal\ \ instructional\ \ leadership) \\ & + \gamma_{01}\,(innovative\ \ teaching\ \ encouragement) \\ & + \mu_{0j}, \mu_{0j} \sim N\,(0, \tau_{00}) \end{aligned} \tag{10}$$

γ_{03}–γ_{07} are the coefficients (effects) of the covariates on the intercept, other parameters have been defined in the baseline model.

4 The Empirical Result Analysis

4.1 Descriptive Statistical Analysis

4.1.1 Descriptive Analysis Statistics of Regional Teachers' Dimensions and Overall Job Satisfaction

Based on the survey sample from 163 districts in 7 cities, we found that the average job satisfaction of elementary school teachers on the 5 dimensions descend from urban, township, to rural areas. Different regions had the same rank orders on satisfaction with respect to interpersonal relationships, leadership and management, environment for development, self-fulfilling, and reasonableness of the effort-reward, from high to low (Table 1). In addition, on the dimension of effort-reward, all regions had the lowest satisfaction. The average score for township and rural area teachers were 2.402 and 2.43, respectively, and it was slightly higher than 2.5 for urban teachers

4.1.2 Descriptive Statistics on Job Satisfaction of Different Teacher Groups

Elementary school TJS differed very slightly across demographic variables. Generally, female teachers had higher satisfaction degrees than their male colleagues. The higher their education background and job tile were, the lower the satisfaction. A weak U-type was found between satisfaction and years of teaching (Table 2).

Table 2 Descriptive statistics of the different characteristics group of teacher job satisfaction

Variables	Dimensions	*n*	*M*	*SD*
Gender	Female	9992	3.671	0.588
	Male	3375	3.625	0.596
Education background	Graduate and above	3487	3.649	0.586
	Undergraduate	9651	3.661	0.592
	Below undergraduate	227	3.706	0.616
Years of teaching	Within 5 years	2050	3.788	0.628
	5–10 years	2051	3.668	0.623
	11–20 years	4956	3.605	0.589
	More than 20 years	4323	3.656	0.548
Job title	Below senior	5387	3.690	0.627
	Senior and above	7600	3.636	0.561

4.2 Correlation Analysis

4.2.1 Correlation Analysis of School Variables and Teacher Job Satisfaction

Table 3 shows the correlations between the school variables and TJS. It appeared that these school variables were moderately correlated with TJS with r between .185 and .423.

4.2.2 Correlation Analysis of Teacher Variables and teachers' job Satisfaction

As shown in Table 4, the teacher variables were moderately correlated with TJS with r between −.148 and .405.

4.3 Analysis on Factors That Influence Elementary School Teacher Job Satisfaction

4.3.1 Differences Between Teacher Job Satisfaction Among Schools

Model_0 (Tables 5 and 6) indicated that school-level random error was 0.261 ($p < 0.001$), which means there was a significant difference in TJS among schools. A total of 24.6 % of the variations came from the school level, which is to say, it was appropriate to apply HLM to study TJS in elementary schools.

4.3.2 Teachers Demographic Variables and School Location Variables Effect on TJS

Model_1 (Tables 5 and 6) shows that: (a) For teacher level variables, in addition to professional titles, gender ($\beta = -0.017$, $SE = 0.011$), education background ($\beta = -0.047$, $SE = 0.012$) and years of teaching ($\beta = -0.081$, $SE = 0.015$) all had a significant impact on TJS; female teachers had a higher job satisfaction than male teachers, the higher the education background and the longer the years of teaching were, the lower the job satisfaction. (b) Geographical location influenced TJS significantly, Rural and township teachers had a significantly lower job satisfaction than urban teachers ($\beta = -0.157$, $SE = 0.034$ and $\beta = -0.073$, $SE = 0.035$, respectively). In short, urban elementary school teachers had the highest job satisfaction, followed by township teachers.

Table 3 Correlations between school variables and teacher job satisfaction

Variables	Democratic decision-making	PD support	Teaching supervision	Principal instructional leadership	Encouragement of teaching innovative	TJS
Democratic decision-making	1.000					
Teachers' professional development support	.790**	1.000				
Teaching supervision	.396**	.410**	1.000			
Principal instructional leadership	.746**	.774**	.483**	1.000		
Encouragement of innovative teaching	.747**	.781**	.446**	.724**	1.000	
TJS	.402**	.423**	.185**	.423**	.370**	1.000

*P < 0.05; **P < 0.01; ***P < 0.001

Table 4 Correlations between teacher variables and teacher job satisfaction

Variables	Daily workload	Monthly income	Part-time situation	Occupational preference	Job involvement	Self-directed PD	TJS
Daily workload	1.000						
Monthly income	.042**	1.000					
Part-time job situation	−.083**	.029**	1.000				
Occupational preference	−.100**	.033**	.017*	1.000			
Job involvement	.031**	.051**	−.064**	.324**	1.000		
Self-directed professional development	.041**	.104**	−.033**	.126**	.346**	1.000	
TJS	−.148**	.056**	.044**	.405**	.376**	.280**	1.000

*P < 0.05; **P < 0.01; ***P < 0.001

Table 5 Fixed effects on TJS in different models

	Model_1		Model_2		Model_3		Model_4	
Fixed effect	Coefficient	S. E	Coefficient	S. E	Coefficient	S. E	Coefficient	S. E
Intercept	13.379***	0.379	14.337***	0.429	9.449***	0.400	2.403***	0.457
Teacher level fixed effects in different models								
Gender	−0.017**	0.011	−0.028***	0.011	0.013**	0.009	0.009*	0.009
Education background	−0.047***	0.012	−0.045***	0.012	−0.030***	0.010	−0.032***	0.010
Years of teaching	−0.081***	0.015	−0.090***	0.015	−0.095***	0.013	−0.090***	0.013
Job title	0.005	0.013	−0.008	0.013	0.001	0.011	0.005	0.011
Daily workload			−0.146***	0.011	−0.137***	0.009	−0.119***	0.009
Monthly income			0.079***	0.014	0.049***	0.012	0.062***	0.010
Part-time situation			0.041***	0.010	0.050***	0.009	0.046***	0.009
Occupational preference					0.286***	0.010	0.275***	0.010
Job involvement					0.250***	0.011	0.223***	0.011
Teacher-based professional development					0.164***	0.010	0.133***	0.010

*P < 0.05; **P < 0.01; ***P < 0.001

Table 6 School level fixed effects in different models

	Model_1		Model_2		Model_3		Model_4	
Rural areas vs. urban	−0.157***	0.034	−0.136***	0.035	−0.035*	0.039	0.017*	0.027
Township vs. urban	−0.073**	0.035	−0.040*	0.036	0.038*	0.038	0.042**	0.027
Democratic decision-making							0.172***	0.045
Teachers' professional development support							0.474***	0.046
Teaching supervision							−0.088***	0.025
Principal instructional leadership							0.384***	0.041
Innovative teaching encouragement							−0.046**	0.044
Random effect in different models								
Between—school level	0.261***	0.005	0.256***	0.005	0.206***	0.004	0.205***	0.004
Within–school level	0.085***	0.005	0.078***	0.004	0.045***	0.003	0.013***	0.002
ICC	0.246		0.233		0.179		0.060	
Between—school variation interpretation ratio		1.92 %		19.53 %		0.49 %		
Within–school variation interpretation ratio		8.24 %		42.31 %		71.11 %		

*P < 0.05; **P < 0.01; ***P < 0.001

4.3.3 Teacher Objective Variables Have Effects on Teacher job Satisfaction

According to Model_2 (Tables 5 and 6), we found: (a) Daily workload ($\beta = -0.146$, $SE = 0.011$), monthly income ($\beta = 0.079$, $SE = 0.014$), and part-time job situation ($\beta = 0.041$, $SE = 0.01$) had a significant effect on job satisfaction. Fewer workloads, fewer part-time job situations, and higher monthly income all resulted in higher job satisfaction. (b) Compared with Model_1, teacher objective variables could explain 1.92 % of job satisfaction variations, and 8.24 % of variations among schools.

4.3.4 Teachers Subjective Variables Effecting on Teachers' Job Satisfaction

According to Model_3 (Tables 5 and 6), the following conclusions were drawn: (a) Occupational preference ($\beta = 0.286$, $SE = 0.01$), job involvement ($\beta = 0.25$, $SE = 0.011$), and self-directed professional development ($\beta = 0.164$, $SE = 0.01$) had a very significant effect on job satisfaction: the higher the career preferences, the higher job satisfaction. When job involvement was higher, job satisfaction was also higher; more professional development initiative led to a higher job satisfaction. (b) Teacher subjective variables could explain 19.53 % of within-school teachers' job satisfaction variations, and 42.31 % of between-school variations.

4.3.5 School Institutional Culture Effect on Teacher Job Satisfaction

From Model_4 (Tables 5 and 6), the following conclusions were drawn: (a) Democratic decision-making ($B = 0.172$, $SE = 0.045$), professional development support ($\beta = 0.474$, $SE = 0.046$), teaching supervision ($\beta = -0.088$, $SE = 0.025$), principal instructional leadership ($\beta = 0.384$, $SE = 0.041$), and teaching innovation encouragement ($\beta = -0.046$, $SE = 0.044$) together had a very significant effect on job satisfaction. The higher the participation in democratic decision-making and principal's leadership, and the more the professional development support, the higher the job satisfaction. (b) Compared with Model_3, school system environment could explain 0.49 % of the variations in job satisfaction and 71.11 % of within-school variations.

5 Discussion and Conclusions

Using the survey data from the CNICA towards Basic Education Quality on regional projects, this study not only investigated urban-rural differences and between-school differences, but also examined to what extent subjective and objective factors of

teachers and school institutional culture variables influenced TJS in elementary schools. The main conclusions and discussion are as follow.

5.1 Differences Among Schools

Differences in TJS among schools are relatively significant. Model_0 shows that 24.7 % of the variations in TJS come from school differences, which is larger than the research by Zhao's 5.41 % (2011) and Shen's 17 % (2012). The main reasons may be: (a) The samples in Zhao's (2011) research are all from the same Anhui province, so the variations were less obvious than this study; (b) There is no clear urban and rural dualistic structure in the United States, so the variations among schools are less obvious than in China. Further studies are needed to examine these explanations.

5.2 Effects of School Location

School location (urban, towns, rural) have significant influences on TJS, with urban the highest, followed by township and rural areas. Urban and rural dualistic structure is seen as the main factor causing the differences, not only in economy, politics, culture, but also in education. There was no large-scale study on urban or rural elementary school teachers' job satisfaction before, and this study was the first one to explore and validate urban-rural differences using a large and representative sample of China. The conclusion of this study is different from the research by Iqbal and Akhtar (2012), who stated that no significant difference was found in TJS between urban and rural secondary school teachers, but is in accordance with the research by Mukhopadhyaya and Kabirak (2014), who found there was a significant difference between urban and rural teachers. Future studies can be conducted to explore differences and similarity in TJS across countries.

5.3 Influential Effects of Teachers' Demographic Variables

Variables such as gender, education background, and years of teaching have a significant influence on TJS, but job title does not. The conclusion on gender is consistent with the study by Shen et.al. (2012) and Crossman and Harris (2006), and the conclusion on years of teaching and education background is consistent with the research by Zhao (2011) and Shen et al. (2012). However, on job title, this study is not in full consistence with the research by Xu and Zhao (2012) and Wu (1996). Further analyses with the survey data from China's National Assessment of Education Quality regional projects might shew some light on this issue.

5.4 *Influence Effects of the Objective Factors of Teachers*

Objective factors of teachers have a significant impact. Teachers who have a part-time job show a lower satisfaction than those who do not have a part-time job. The higher the monthly income is, the higher job satisfaction; and the higher the daily workload is, the lower satisfaction. This conclusion is very similar to the research by Chen and Sun (1994), but different from the study by Korb and Akintunde (2013). The main cause may be that currently in China, in addition to daily teaching, elementary school teachers have extra administration work to do. The study showed that 56.7 % of teachers have additional workload, 44 % have a workload of more than 10 h per day, but their income does not match their workloads. The survey shows that effort-reward reasonableness hit the lowest in all dimensions around all areas (Table 1). For monthly income, 34.9 % of teachers earn less than 3000 yuan, 30.1 % reach to 3000–4000. The greatest pressure comes from the work load factors (including long working hours, heavy workload, and demanding work expectations), followed by job guarantee factors and teaching security factors. Therefore, the present study agrees with the conclusion in Prick (1989) that work stress has a significantly negative prediction on job satisfaction, and together with job satisfaction degree could have direct and indirect influence on job burnout.

5.5 *Influence Effects of Teachers' Subjective Factors*

Teachers' subjective factors have a significant impact on TJS, and occupational preferences as well as job involvement have very evidential prediction to TJS (prediction coefficients are 0.286 and 0.25, both with $p < 0.000$; Tables 5 and 6). This conclusion is basically in consistence with the studies by Li (2013), Hackman and Oldham (1976). This study also found that teachers' subjective factors have higher impacts on job satisfaction than the objective factors do, and job satisfaction is higher when occupational preference and job involvement are higher, which is consistent with John Holland's vocational interest theory. On the other hand, according to Maslow's hierarchy of needs theory that self-realization ranks the highest, teachers' aspirations for self-realization are higher than general public because teachers are communicators of human culture and scientific knowledge. Our results support past research in that self-efficacy is linked to job satisfaction. Therefore, it is not surprising that elementary school teachers have high professional development needs and such needs have a significant positive effect of their job satisfaction.

5.6 Influence Effects of School System Environment

Democratic decision-making, professional development support, and principal leadership have a positive prediction and significant effect on TJS, while teaching supervision and teaching innovation encouragement have an obvious negative prediction. In general, the school system is the unity of the willingness of school and teachers, so the development of schools and achievement of educational objectives call for a fair and equitable system and charisma of principal and teacher self-development consciousness. Our results support past research in that decision-making, professional development support, and principal's instructional leadership are linked to job satisfaction. Li (2013) indicated that the school system environmental factors have a significant impact on job satisfaction of beginning teachers, and to raise TJS attention needs to focus on the key role that school system factors play, as well as the beginning teachers' positive professional values and a favorable environment for their growth. Aldridge and Fraser (2015) agreed that approachable and supportive school principals contribute both directly and indirectly to TJS. Olcum and Titrek (2015) suggested that degree of TJS can be predicted significantly by administrators' decision-making styles. Dadkhah, Radzi, Huang, and Jenatabadi (2014) indicated that transformational leadership and participative decision making have a significant impact on TJS in universities. Shen et al. (2012) recommended that school principals had a significant effect on TJS. All these studies indicate that the effects of the school system environment on TJS in elementary schools are very significant. In the present study, the explanation percentage in Model_4 is improved by 71.6 %, as compared to Model_3. Therefore, it can be concluded that although the variations in TJS is mainly attributable to individual teachers, school can play a very important role to improve job satisfaction.

6 Suggestions

Based on the findings, we make the following suggestions. In order to improve teachers' job satisfaction in China, school leaders should develop a positive environment, increase teachers' participation in democratic decision-making, and strengthen teachers' professional development support. Education authorities should provide effective trainings on instructional leadership to school principals, reduce teachers' workloads, and raise their salary to boost teachers' job satisfaction, which in turn will improve the quality of education.

References

Aldridge, J. M., & Fraser, B. J. (2015). Teachers' views of their school climate and its relationship with teacher self-efficacy and job satisfaction. *Learning Environments Research*. doi: 10.1007/s10984-015-9198-x.

Arifin, H. M. (2015). The influence of competence, motivation, and organisational culture to high school teacher job satisfaction and performance. *International Education Studies, 8*(1), 38–45. doi:10.5539/ies.v8n1p38.

Bogler, R. (2002). Two profiles of schoolteachers: A discriminant analysis of job satisfaction. *Teaching and Teacher Education, 18*(6), 665–673. doi:10.1016/S0742-051X(02)00026-4.

Caprara, G. V., Barbaranelli, C., Steca, P., & Malone, P. S. (2006). Teachers' self-efficacy beliefs as determinants of job satisfaction and students' academic achievement: A study at the school level. *Journal of School Psychology, 44*(6), 473–490. doi:10.1016/j.jsp.2006.09.001.

Chen, Y., & Sun, S. (1994). A measurement study of teachers' job satisfaction. *Psychological Science, 17*(3), 146–150. doi:10.1017/CBO9781107415324.004.

Crossman, A., & Harris, P. (2006). Job satisfaction of secondary school teachers. *Educational Management Administration & Leadership, 34*(1), 29–46. doi:10.1177/1741143206059538.

Dadkhah, V., Radzi, C. W., Huang, H., & Jenatabadi, H. S. (2014). School size and the relationship among principal's leadership style, decision making style, and teacher job satisfaction: A moderation analysis. *International Journal of Research in Business and Technology, 5*(3), 724–729. doi:10.2139/ssrn.2383912.

Feng, B. (1996). Study on teachers' job satisfaction and its influencing factors. *Journal of Education Research, 2*, 42–49.

Friedlander, F. (1964). Job characteristics as satisfiers and dissatisfiers. *Journal of Applied Psychology, 48*(6), 388.

Hackman, J. R., & Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance, 16*(2), 250–279. doi:10.1016/0030-5073(76)90016-7.

Herzberg, F. (1968). One more time: How do you motivate employees? *Harvard Business Review*. doi:10.1108/eb055227.

Hoppock, R. (1935). *Job satisfaction*. Oxford, England: Harper.

Hu, Y. (2007). A survey report on junior middle school teachers' job satisfaction and influence factors. *Journal of Education, 3*(5), 46–52.

Iqbal, A., & Akhtar, S. (2012). Job satisfaction of secondary school teachers. *Abasyn Journal of Social Sciences, 5*(1), 49–65.

Judge, T., Judge, T., Heller, D., Heller, D., Mount, M., & Mount, M. (2002). Five-factor model of personality and job satisfaction. *Journal of Applied Psychology, 87*(3), 530–541. doi:10.1037/0021-9010.87.3.530.

Klassen, R. M., & Tze, V. M. C. (2014). Teachers' self-efficacy, personality, and teaching effectiveness: A meta-analysis. *Educational Research Review, 12*, 59–76. doi:10.1016/j.edurev.2014.06.001.

Korb, K. A., & Akintunde, O. O. (2013). Exploring factors influencing teacher job satisfaction in Nigerian schools. *Nigerian Journal of Teacher Education and Training*, *11*, 211–223. Retrieved from http://korbedpsych.com/LinkedFiles/Nigerian_Teacher_Job_Satisfaction.pdf.

Li, M. (2013). A research on influencial dactors of lower school new teachers' job satisfaction. *Teacher Education Research, 25*(5), 43–48. doi:10.13445/j.cnki.t.e.r.2013.05.011.

Liu, X. S., & Meyer, J. P. (2005). Teachers' percetions of their jobs: A multilevel analysis of the teacher follow-up survey for 1994–95. *Teachers College Record, 107*(5), 985–1003. doi:10.1111/j.1467-9620.2005.00501.x.

Locke, E. A. (1976). The nature and causes of job satisfaction. *Handbook of Industrial and Organizational psychology, 1*, 1297–1343.

Luthans, F., Zhu, W., & Avolio, B. J. (2006). The impact of efficacy on work attitudes across cultures. *Journal of World Business, 41*(2), 121–132. doi:10.1016/j.jwb.2005.09.003.

Miao, N. (2006). On the function of the teacher in charge of a class in quality education of higher vocational colleges. *Journal of Changzhou Vocational College of Information Technology, 2*, 85–88.

Mukhopadhyaya, D., & Kabirak, U. S. (2014). Educational research department of education university of Calcutta. *Indian Journal of Educational Research, III*, 200–207.

Olcum, D., & Titrek, O. (2015). The effect of school administrators' decision-making styles on teacher job satisfaction. *Procedia—Social and Behavioral Sciences, 197*(February), 1936–1946. doi:10.1016/j.sbspro.2015.07.575.

Prick, L. G. (1989). Satisfaction and stress among teachers. *International Journal of Educational Research, 13*(4), 363–377.

Schaufeli, W. B., & Bakker, A. B. (2003, November). UWES Utrecht Work Engagement Scale Preliminarty Manual. *Journal of Occupational Health Psychology*, *58*. doi:10.1037/t01350-000.

Shen, J., Leslie, J. M., Spybrook, J. K., & Ma, X. (2012). Are principal background and school processes related to teacher job satisfaction? A multilevel study using schools and staffing survey 2003–04. *American Educational Research Journal, 49*(2), 200–230. doi:10.3102/0002831211419949.

Skaalvik, E. M., & Skaalvik, S. (2010). Teacher self-efficacy and teacher burnout: A study of relations. *Teaching and Teacher Education, 26*(4), 1059–1069. doi:10.1016/j.tate.2009.11.001.

Veldman, I., van Tartwijk, J., Brekelmans, M., & Wubbels, T. (2013). Job satisfaction and teacher-student relationships across the teaching career: Four case studies. *Teaching and Teacher Education, 32*, 55–65. doi:10.1016/j.tate.2013.01.005.

Wang, H., Hall, N. C., & Rahimi, S. (2015). Self-efficacy and causal attributions in teachers: Effects on burnout, job satisfaction, illness, and quitting intentions. *Teaching and Teacher Education, 47*, 120–130. doi:10.1016/j.tate.2014.12.005.

Wu, K. F. J. (1996). Job satisfaction of Hong Kong secondary school teachers. *Education Journal, 24*(2), 29–44.

Xu, Z., & Zhao, Z. (2012). An empirical study of job satisfaction of Beijing primary school teachers. *Teacher Education Research, 24*(1), 85–92.

Zembylas, M., & Papanastasiou, E. (2004). Job satisfaction among school teachers in Cyprus. *Journal of Educational Administration, 42*(3), 357–374. doi:10.1108/09578230410534676.

Zhao, B. (2011). The factors which influenced teachers' job satisfaction: Base on analysis of the HLM. *Education Science, 27*(4), 30–34.

The Determinants of Training Participation, a Multilevel Approach: Evidence from PIAAC

Teck Kiang Tan, Catherine Ramos, Yee Zher Sheng, and Johnny Sung

Abstract This chapter uses the first round of the Programme for the International Assessment of Adult Competencies (PIAAC) survey data to find out the determinants of training participation of workers for the 24 countries that participated in the survey. Two measures are used in quantifying training participation: (1) whether workers participated in any training during the last 12 months and (2) the number of training modes of participation (number of types of trainings workers attended during the last 12 months). Logistic and Poisson multilevel models were used to model the two measures respectively. Both models show similar findings for the fixed effects. The results of the random slopes models show that heterogeneity exists across the 24 countries, indicating that the effects of covariates on workers' training participation and the number of training modes vary across countries. The magnitude of correlations of these random slopes differs between the logistic and Poisson models, indicating that the associations of these random effects are not totally in agreement between training participation and the number of training modes.

Keywords Training participation • Random slope models • Multilevel models

1 Introduction

Engaging in lifelong learning has been acknowledged as a good attribute that workers should be exercising given the individual and social challenges that adults may face in the current globalized world where speed of technology and ease of communication has been accelerating both at work and outside of work.

T.K. Tan (✉)
Research and Innovation Division, Institute for Adult Learning, 1 Kay Siang Road, Tower Block Level 6, Singapore 248922, Singapore
e-mail: tan_teck_kiang@ial.edu.sg

C. Ramos
The Head Foundation, Singapore, Singapore

Y.Z. Sheng • J. Sung
Institute for Adult Learning, Singapore, Singapore

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_27

For example, the Scottish government acknowledges this importance through the following statements "A skilled and educated workforce is essential to productivity and sustainable economic growth. Not only are more skilled workers potentially more productive in their own right, but the skill level of the workforce is likely to impact significantly on the effectiveness of capital investment and the ability of employers to adopt innovative work practices" (Scottish Government, 2007, p 6, as cited in Sutherland, n.d.).

The recent adult skills survey that was released by OECD (2013) re-emphasized that adult learning can play an important role in developing and maintaining key information skills and in acquiring other knowledge and skills which are necessary to keep pace with the changing work environment. In order to enhance adult training participation, policymakers need to understand the current patterns of training activity as well as the incentives to make individuals pursue training.

With this backdrop, who participates in training, and what do international data from adult skills study tell us? From the inclusive-growth point of view, it is ideal if the vulnerable workers, e.g., workers with less than tertiary educational qualification, have higher incidence of training. But much empirical evidence suggests that high education level is positively correlated with current learning or continuous training (OECD 2013; Biagetti & Scicchitano 2009; Jenkins et al. 2002). Biagetti and Scicchitano (2009) reported that there are few countries in Europe where the engagement in adult education was higher among less educated workers—only Finland, Latvia, Denmark for men, and only Finland, Hungary, Lithuania, and UK for women. The researchers noted that Denmark is the only country where being less educated is the most relevant variable for engaging in lifelong learning. The UK Commission for Employment and Skills Report Johnson et al. (2009) on review of evidence on factors affecting participation in skills development noted that evidence appears to suggest "that individuals being higher-qualified and higher skilled is a crucial predictor of access to work-related training; i.e., levels of access appear to be highest among those who are least disadvantaged" (p. 18). Other determinants of training were also identified in terms of personal (e.g., age, family responsibility), employer (e.g. firm size, industry), and job-related (e.g., occupational type, job status) characteristics. For example, the study by Fritsche (2012) on determinants of training participation showed that training participation increases with age up to certain point where the probability of participation started to decrease, implying an inverted U-shape relationship between training participation and age. His research also confirms the many findings about the positive and significant relationship of education level and training participation. Being female also yielded a higher likelihood for training participation according to the study. In the same study, income, tenure, being a civil servant, being in a larger company and following a healthy diet all has positive and significant effects on training participation. The OECD (2003) has summarized some issues related to adult learning. For example, participation is highly unequal between certain groups, i.e., younger adults, and those with higher educational attainment which workers in high-skilled occupations tend to have more access to learning opportunities than others.

While research on determinants of training participation tends to focus more on the socio-education-economic variables (such as age, gender, education, family background, employment status, income), this chapter's contributes to the literature by looking at how skills utilization affects continuing education and training. Skills utilization has become popular and important in the skills discourse and it has been argued that skills utilization is just as important as educational attainment and investment.

Multilevel models with random intercepts and slopes are used in this chapter to examine the effect of skills utilization on continuing training and education. Such models are commonly used in research among subjects within the same ecological environments that are consequently correlated (Bryk & Raudenbush 2002; Bickel 2007). For instance, worker participation in training has a more similar social and economic environment for workers in the same country compared to workers in different countries. The use of multilevel models for analyses has become common in such contexts as it adjusts for the biased standard errors for covariates effects in a single regression model. Most often, the random intercept model is used in academic disciplines (e.g. Berigan & Irwin 2011; Correnti, Matsumura, Hamilton & Wang 2013; Curci, Lanciano & Soleti 2014; Lim & MacGregor 2012) but the more complex random slope models are often ignored probably due to lack of understanding of the concept and use of heterogeneity in answering relevant research functions. The omission of random slopes model is not always inappropriate as the heterogeneity of higher level covariates is of research interest and the data support it. This chapter illustrates the use of random slope models to examine the heterogeneity between countries and argues that methodologically random slope models should be used in order to examine and understand variation in covariate effects at the country level.

2 Measures and Source of Data

The data were extracted from the first round of PIAAC for the 24 countries participating in the survey. The analysis is restricted to respondents who were working during the survey period so that the explanatory variables such as literacy and numeracy used at work are applicable for the current study.

During the survey, workers were asked for the period of the last 12 months whether they had participated in courses conducted through open or distance learning, attended organized sessions for on-the-job training or training by supervisors or co-workers, participated in seminars or workshops, participated in courses or private lessons, or were receiving a formal education. These modes of learning could take place either at work or within a personal capacity. These five modes of training are summarized into two measures of training participation. The first measure quantifies whether workers participated in training. It is coded as one if workers participated in any of the five modes of training, otherwise it is coded as zero. The second measure determines the number of training modes by counting number of occurrences across

the five modes of training. If workers participated in all five modes of training, the variable is coded as five, and if workers did not participate in any, it is coded as zero. A logistic model is used to analyse the former outcome while a Poisson model for the latter outcome.

Our model concentrates on examining whether the frequency of numeracy and literacy used at work and in daily life affects the training participation of workers. The literacy variable captures skills use with respect to writing letters, memos, mails, reports and filling in forms. The numeracy variable includes calculating costs or budgets, using fractions, percentages, using a calculator at work, preparing charts, graphs or tables, using simple algebra, formulas or advanced mathematics or statistics. The literacy and numeracy skills variable are principal component scores based on three and six indicators, respectively.

Six control variables are included in the modelling. These include gender, age, highest education level of workers, socio-economic status of parents, industry, and firm size. For comparison across countries, education level and industry are coded using the International Standard Classification of Education (ISCED) and International Standard Industry Classification (ISIC), respectively. These two codes are further regrouped into four and ten categories respectively for education level and industry. Gender and highest education level are coded as dummy variables. The reference group for gender is male. For education level, the reference category is "lower secondary education and below". Parent's socio-economic status (SES) refers to father's and mother's highest educational level.

3 Description of Models

The multilevel logistic model is specified as follows:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{u}_j + e_{ij},$$
$$p_{ij} = P\left(Y_{ij} = 1\right),$$

where $\mathbf{X\beta}$ consists of the fixed components and $\mathbf{Zu}$ consists of the random components.

$$\mathrm{i} = 1, \ldots, 112,651 \text{ workers},$$
$$\mathrm{j} = 1, \ldots, 24 \text{ countries, and}$$

$e_{ij} \sim N\left(0, \sigma_e^2\right)$.

Twelve covariates are specified as having random slopes. They are educational level (with three dummies), age group (with four dummies), socio-economic status, literacy use at work, literacy use for daily life, numeracy use at work, and numeracy use for daily life. The matrix representation of the 12 random slopes is specified below.

$$\mathbf{U}_j = \begin{pmatrix} u_{oj} \\ u_{1j} \\ \vdots \\ u_{11j} \\ u_{12j} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} \sigma^2_{u0} & \sigma_{0,1} & \dots & \sigma_{0,11} & \sigma_{0,12} \\ \sigma_{0,1} & \sigma^2_{u1} & \dots & \sigma_{1,11} & \sigma_{1,12} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_{0,11} & \sigma_{1,11} & \dots & \sigma^2_{u11} & \sigma_{11,12} \\ \sigma_{0,12} & \sigma_{1,12} & \cdots & \sigma_{11,12} & \sigma^2_{u12} \end{pmatrix} \right]$$

The multilevel Poisson model is specified below. Similar to logistic model, the same sets of covariates together with the 12 random slopes are included in the Poisson model.

Specifically,

$$\begin{aligned} n_{ij} &\sim \text{Poisson}\left(\lambda_{ij}\right), \\ \log \lambda_{ij} &= \eta = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}_{ij}, \end{aligned}$$

$$\begin{aligned} &\text{i} = 1, \dots, 112,651 \text{ workers} \\ &\text{j} = 1, \dots, 24 \text{ countries}, \end{aligned}$$

$$e_{ij} \sim N\left(0, \sigma^2_e\right), \text{ and}$$

$$\mathbf{U}_j = \begin{pmatrix} u_{oj} \\ u_{1j} \\ \vdots \\ u_{11j} \\ u_{12j} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} \sigma^2_{u0} & \sigma_{0,1} & \dots & \sigma_{0,11} & \sigma_{0,12} \\ \sigma_{0,1} & \sigma^2_{u1} & \dots & \sigma_{1,11} & \sigma_{1,12} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_{0,11} & \sigma_{1,11} & \dots & \sigma^2_{u11} & \sigma_{11,12} \\ \sigma_{0,12} & \sigma_{1,12} & \cdots & \sigma_{11,12} & \sigma^2_{u12} \end{pmatrix} \right].$$

Exploratory factor analyses were carried out for the four skills variables, namely literacy skills use at work, literacy skills use in daily life, numeracy skills use at work, and numeracy skills use in daily life. These four skills variables are standardized to mean zero and standard deviation one. Socio-economic status is a principal component score of father's and mother's highest educational level with mean zero and standard deviation one. Package R lme4 function glmer is used to fit multilevel logistic and Poisson models.

4 Results

The descriptive statistics for the variables are listed in Table 1. Overall, about 69 % of workers participated in training across the 24 countries. On average, they participated in 1.01 training modes out of the 5 modes of training. About 2 out of 3 workers (57 %) worked in a firm having less than 50 employees and about 2 out of 5 (41 %) attained qualification with a diploma or degree. About half of the respondents were female (49 %), and 70 % were between the ages of 25–54, and 1 in 3 (34 %) worked in the community, social, or personal services.

Table 1 Descriptive statistics

Variable	Metric	Mean	SD
Participation rate	Range from 0 to 1	0.69	0.48
Number of participation		1.01	1.04
Skills use			
Literacy at work	Standardized score	0	1
Literacy for daily life	Standardized score	0	1
Numeracy at work	Standardized score	0	1
Numeracy for daily life	Standardized score	0	1
Demographics/control variables			
Gender	Male = 0, Female = 1	0.49	–
Age group			
Below 25	(omitted)	0.13	–
25–34	25–34 = 1	0.22	–
35–44	35–44 = 1	0.24	–
45–54	45–54 = 1	0.24	–
55 and above	55 and above = 1	0.17	–
Education level			
Primary and below	(omitted)	0.14	–
Post Secondary	Post secondary = 1	0.45	–
Diploma	Diploma = 1	0.12	–
Degree	Degree = 1	0.29	–
Socio-economic status of parents	Standardized score	0	1
Industry			
Manufacturing	(omitted)	0.13	–
Construction	Construction = 1	0.07	–
Wholesale and retail trade	Wholesale and retail trade = 1	0.14	–
Transportation and storage	Transportation and storage = 1	0.05	–
Accommodation and food services	Accommodation and food = 1	0.05	–
Information and communications	Information and comm. = 1	0.03	–
Financial and insurance services	Financial and insurance = 1	0.03	–
Business services	Business services = 1	0.10	–
Community, social and personal services	Community and services = 1	0.34	–
Others	Others = 1	0.05	–
Firm size			
10 and below	(omitted)	0.26	–
11–50	11–50 = 1	0.31	–
51–250	51–250 = 1	0.23	–
251–1000	251–1000 = 1	0.12	–
More than 1000	More than 1000 = 1	0.08	–

5 Fixed Effects

Most of the results of fixed effects for the Poisson and logistic models show similar findings. Specifically, literacy and numeracy skills use at work and in everyday life affect the training participation of workers. Younger workers are more likely to participate in training than older workers. Workers with higher education and having parents with higher education are more likely to participate in training than less educated workers having parents with less education. Workers in larger firms, and with higher literacy use at work, literacy use in everyday life, and numeracy use in everyday life are more likely to participate in training. The results also reveal the differential participation of workers between industries. Those working in information and communications, finance, and insurance, business services, community, social and personal services have higher likelihood in training participation as compared to those working in other industries.

The gender effect differs for the two models. The Poisson model shows that females are statistically higher in the number of training modes whereas the logistic model shows the insignificant difference between genders in participation. The level of significance also shows differences for the two models such as numeracy at work (Table 2).

6 Random Effects

While the fixed effects measure the overall mean effects of the covariates on training participation, the random effects quantify the magnitude of variation of the estimated covariates. In general, the higher the variance of the random slope estimates, the greater the variation and vice versa. For testing the significance of these random effects, deviance tests for both logistic and Poisson models are carried out. The results show highly significant with Chi-Square values of 256.71 and 309.23 respectively for logistic and Poisson models. Consistent results for AICs also show a drop from 76,911 to 76,834, and 168,775 to 168,646 respectively for the logistic and Poisson models.

Figure 1 shows the estimates of the posterior modes of the random slope (Bates & Bolker 2015) for the literacy use at work across the 24 countries for the Poisson model. The estimated slopes for the 24 countries range from 0.192 (USA) to 0.351 (France), providing evidence of heterogeneity across the 24 countries. Taking the top two highest and lowest countries, Italy (0.333) is not much different from France (0.351), and Norway (0.212) is close to the USA (0.192), but the magnitudes of literacy at work in training participation for the highest two countries do differ substantially from the lowest two countries. The random slope model shows the variation of literacy at work for the 24 countries. If the modelling is restricted to a basic random intercept model, these variations will be ignored, providing only a fixed point estimate for all the countries, disregarding the variation in effects of the covariate.

Table 2 Results from multilevel logistic and Poisson models

Variable	Poisson	Logistic
Fixed effect		
Intercept	−0.38***	0.72***
Female	0.05***	0.03
Age (ref: below 25)		
25–34	−0.32***	−1.04***
35–44	−0.36***	−1.12***
45–54	−0.37***	−1.10***
55 and above	−0.51***	−1.39***
Education level (ref: primary and below)		
Post secondary	0.24***	0.11*
Diploma	0.35***	0.35***
Degree	0.40***	0.43***
Industry (ref: manufacturing)		
Construction	−0.06**	0.00
Wholesale and retail trade	0.03	0.07*
Transportation and storage	0.04	0.10*
Accommodation and food services	0.04	0.12*
Information and communications	0.10***	0.13*
Financial and insurance services	0.27***	0.65***
Business services	0.10***	0.17***
Community, social and personal services	0.25***	0.58***
Others	0.16***	0.32***
Firm size (ref: 10 and below)		
11–50	0.16***	0.33***
51–250	0.25***	0.57***
251–1000	0.31***	0.74***
More than 1000	0.34***	0.82***
Socio-economic status	0.04***	0.09***
Literacy at work	0.27***	0.53***
Literacy for daily life	0.15***	0.23***
Numeracy at work	−0.00	−0.04*
Numeracy for daily life	0.03***	0.15***

***p < .001, **p < .01, *p < .05

While the variances of the random slopes measure the extent of variation, the correlations among the random slopes measure the extent of association of the covariate effects at the country level. Comparing the differences in the magnitude of correlations between the logistic and Poisson models also reveals whether the effects on participation and the effects on the number of modes respondents chosen show similar relationships.

Table 3 and Table 4 show the estimated correlations of the random slopes for the 12 covariates for the logistic and Poisson models respectively. The correlation

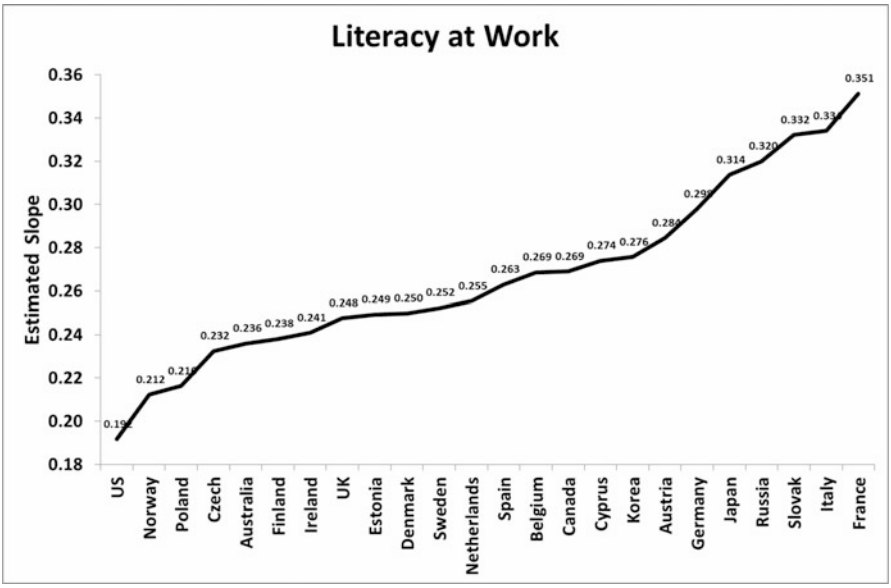


Fig. 1 Heterogeneity of literacy at work—Poisson model

between the random slopes of literacy use in daily life and literacy use at work is negative with a high value of −0.78 for the logistic model (Table 3) but there is hardly any association (0.08) for the Poisson model (Table 4). The negative correlation for the logistic model indicates that countries having lower coefficients for literacy use at work would tend to have higher coefficients for literacy use in daily life and vice versa. The low positive correlation for the Poisson model tells a different story suggesting there is no such association in relation to the number of training modes. Since the dependent variable of the logistic model is about participation or non-participation in training, while the Poisson model is about the number of training modes that workers participated in, thus, the correlation between literary use (both at work and in every a life) and training participation may differ and has actually differed in this study from the correlation between the literary use (both at work and in every a life) and training modes.

7 Conclusion

This chapter examines the determinants of training participation using two random intercept and slope models, and using data for the 24 countries PIAAC-participating countries. Most of the results for fixed effect show that both logistic and Poisson multilevel models have similar findings in that socio-demographic, employment-related characteristics, the skills used at work and in everyday life, do affect

Table 3 Correlation of random slopes, multilevel logistic model

	Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
(1)	Intercept												
(2)	Post Secondary	−0.66											
(3)	Diploma	−0.10	0.48										
(4)	Degree	−0.29	0.70	0.65									
(5)	Literacy at work	0.49	−0.11	0.36	0.16								
(6)	Literacy for daily life	−0.19	−0.14	−0.71	−0.25	−0.78							
(7)	Numeracy at work	−0.49	0.06	−0.45	0.02	−0.13	0.29						
(8)	Numeracy for daily life	0.73	−0.85	−0.26	−0.62	0.40	−0.21	−0.36					
(9)	Socio-economic status	−0.06	−0.06	−0.11	−0.32	−0.34	0.14	−0.54	0.35				
(10)	Age 25–34	−0.56	0.24	−0.16	−0.02	−0.18	−0.11	0.66	−0.30	−0.14			
(11)	Age 35–44	−0.59	0.35	−0.01	0.20	−0.16	−0.16	0.58	−0.35	−0.07	0.96		
(12)	Age 45–54	−0.60	0.33	−0.05	0.16	−0.24	−0.10	0.48	−0.27	0.15	0.90	0.97	
(13)	Age 55 and above	−0.43	0.18	−0.11	0.25	−0.38	0.13	0.31	−0.16	0.31	0.63	0.78	0.88

Table 4 Correlation of random slopes, multilevel Poisson model

	Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
(1)	Intercept												
(2)	Post secondary	−1.00											
(3)	Diploma	−0.83	0.82										
(4)	Degree	−0.83	0.82	0.85									
(5)	Literacy at work	0.47	−0.44	−0.61	−0.44								
(6)	Literacy for daily life	−0.21	0.21	−0.27	0.10	0.08							
(7)	Numeracy at work	0.55	−0.56	−0.16	−0.20	0.00	−0.27						
(8)	Numeracy for daily life	0.34	−0.37	−0.10	−0.41	−0.61	−0.30	0.36					
(9)	Socio-economic status	0.40	−0.40	−0.47	−0.19	0.22	0.38	−0.02	−0.06				
(10)	Age 25–34	−0.25	0.24	0.57	0.20	−0.63	−0.65	0.19	0.52	−0.68			
(11)	Age 35–44	−0.09	0.07	0.51	0.19	−0.59	−0.73	0.37	0.55	−0.52	0.96		
(12)	Age 45–54	0.04	−0.06	0.36	0.17	−0.57	−0.66	0.34	0.55	−0.29	0.84	0.94	
(13)	Age 55 and above	−0.06	0.03	0.40	0.29	−0.71	−0.47	0.30	0.56	−0.26	0.76	0.85	0.94

training participation and the number of training modes that workers participated in. Younger workers, with higher family SES, working in larger firms, and in areas of information & communications, financial & insurance, business services, and community, social & personal services are more likely to participate in training and be involved in more modes of trainings. These results are mostly consistent with the literature. The random slope models further reveal that heterogeneity exists across the 24 countries and that the associations among these slopes differ between logistic and Poisson models.

Substantial empirical research using multilevel models generally concentrates on examining the fixed effects, not the random effects. Random intercept and slope models, which aim to examine level 2 group differences, are not the norm in research. Crucial information will be left out if research considers only random intercept model. The basic set up of the random slope model emphasizes differences across level 2, in this chapter, the countries, due to heterogeneity. Using random slope models, this chapter managed to examine the heterogeneity for 12 relevant covariates across 24 countries. The incremental information in examining the differential effects of covariates is key in cross-country comparative studies as well as for research agendas with aims to better understand higher levels differences in terms of its variation and covariation. This chapter, hopefully, encourages for future research exploring this useful form of multilevel modelling.

Acknowledgements The research was supported by the Institute for Adult Learning, Research and Innovation Department. We thank Emily Low, Soo Kheng Sim, and administrative staff Zach Aw, Aggie Choo, and Eric Lee for their support to make this chapter possible. We also like to thank Dennis Kwek and Daniel Bolt for their comments and feedback to sharpen and improve the chapter.

References

Bates, D., & Bolker, B. (2015). *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-10*. Retrieved from http://CRAN.R-project.org/.

Berigan, N., & Irwin, K. (2011). Culture, cooperation, and the general welfare. *Social Psyhological Quarterly, 74*, 341–360.

Biagetti, M., & Scicchitano, S. (2009). *What does affect inequality in workers' formal lifelong learning across European countries?* Preliminary draft, version 16 June 2009.

Bickel, R. (2007). *Multilevel analysis for applied research: It's just regression!* New York: Guilford Press.

Bryk, A. S., & Raudenbush, S. W. (2002). *Hierarchical linear models*. Newbury Park, CA: Sage.

Correnti, R., Matsumura, L. C., Hamilton, L., & Wang, E. (2013). Assessing students' skills at writing analytically in response to texts. *The Elementary School Journal, 114*, 142–177.

Curci, A., Lanciano, T., & Soleti, E. (2014). Emotions in the classroom: The role of teachers' emotional intelligence ability in predicting students' achievement. *The American Journal of Psychology, 127*, 431–445.

Fritsche, S. (2012). *Determinants of training participation: A literature review and empirical support from personal-, job-, employer- and health related factors in Germany*. Retrieved June 22, 2014 from http://arno.uvt.nl/show.cgi?fid=128033.

Jenkins, A., Vignoles A., & Galindo-Ruedo, F. (2002). The determinants and labor market effects of lifelong learning. Retrieved from http://eprints.lse.ac.uk/19532/1/The-Determinants-and-Effects-of-Lifelong-Learning.pdf.

Johnson, S., Sawicki, S., Pearson, C., Lindsay, C., McQuaid, R. W., Dutton, M. (2009). *Employee demand for skills: A review of evidence and policy.* Evidence report 3, A Report to UK Commission for Employment and Skills. Retrieved from http://dera.ioe.ac.uk/9116/2/Evidence_Report_3_3.pdf.

OECD. (2003). *Beyond rhetoric: Adult learning policies and practices, highlights*. Retrieved from http://www.oecd.org/education/innovation-education/18466358.Pdf.

OECD. (2013). OECD skills outlook 2013: First results from the survey of adult skills. Retrieved May 22, 2014 from http://www.oecd.org/site/piaac/Skills%20volume%201%20%28eng%29--full%20v12--eBook%20%2804%2011%202013%29.pdf.

Lim, C., & MacGregor, C. A. (2012). Religion and volunteering in context disentangling the contextual effects of religion on voluntary behaviour. *American Sociological Review, 77*, 747–779.

Latent Transition Analysis for Program Evaluation with Multivariate Longitudinal Outcomes

Depeng Jiang, Rob Santos, Teresa Mayer, and Leanne Boyd

Abstract Evaluations of intervention programs, such as the PAX Good Behaviour Game (PAX) program often have multiple outcome variables (e.g., emotional symptoms, conduct problems, hyperactivity/inattention, peer relationship problems, and prosocial behaviour). These are often reported for multiple time points (e.g., pre- and post-intervention) where data are multilevel (e.g., students nested in schools). In this paper, we use latent transition analysis (LTA), a *person-oriented* statistical approach, to evaluate the PAX program with multilevel, longitudinal multivariate outcomes. Using data from the Manitoba PAX Study, we show how LTA helps explore the transition of multiple outcomes across multiple time points and how the intervention program affects this transition. The strengths and limitations of LTA are discussed.

Keywords Intervention • Program evaluation • Latent transition analysis • Longitudinal multivariate outcomes

Evaluations of intervention programs often have multiple outcome variables (e.g., emotional symptoms, conduct problems, hyperactivity/inattention, peer relationship problems, and prosocial behaviour). There is often substantial variability among program participants in terms of intervention effectiveness. Jiang, Pepler and Yao (2010) illustrated the necessity of identifying participant heterogeneity in the design and evaluation of intervention studies, where it is likely that distinct subgroups exhibit different treatment response patterns. Such heterogeneity is often overlooked in the analysis of intervention data, because these data are typically analyzed using *variable-oriented* statistical approaches such as regression. These approaches estimate how the intervention groups differ on outcomes and the results are often presented as an average effect. This may be misleading, because averages may

D. Jiang (✉) • R. Santos
Department of Community Health Sciences, University of Manitoba, Bannatyne Ave, Winnipeg, MB, Canada R3E 0W3
e-mail: depeng.jiang@umanitoba.ca

R. Santos • T. Mayer • L. Boyd
Healthy Child Manitoba Office, Winnipeg, MB, Canada

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_28

represent a mixture of benefit some and harm some. If only a small fraction of participants show the intended intervention outcomes, then *variable-oriented* approaches may fail to detect these effects (Thompson, Mary & Fraser 2011).

Person-oriented statistical approaches focus on individuals, with a goal to classify them into distinct groups (classes, clusters, categories or profiles) based on individual response patterns. Individuals within the same group are similar to each other and different from those in other groups. *Person-oriented* methods of analyzing intervention data offer new directions for investigating developmental issues and provide useful and powerful tools to assess change patterns at individual and group levels.

Latent class analysis (LCA) is one of the most commonly used *person-oriented* approaches. LCA is related to factor analysis, in which the covariation of observed variables is explained by latent continuous variables (factors). LCA differs from factor analysis in that the latter decomposes covariances to highlight relationships among variables, whereas LCA decomposes covariances to highlight relationships among individuals (Bauer & Curran 2004). LCA is similar to traditional cluster analysis, but offers several advantages. Whereas cluster analysis is an exploratory technique, LCA is a model-based procedure that allows for more flexible model specification. The fit indexes provided in LCA enable different models to be compared and inform decisions regarding the number of underlying classes (Pastor, Barron, Miller & Davis 2007).

Latent transition analysis (LTA) is a variant of LCA used for modeling change over time in a discrete developmental process (Collins & Wugalter 1992). Typically, LCA uses manifest indicators that were all measured at the same time. LTA extends LCA to longitudinal data by integrating autoregressive modeling to examine how group membership changes over time. Transitions between profiles from one time point to the next are estimated dependent on baseline profile membership, covariates, and treatment assignment (Thompson et al. 2011).

To date, the application of LTA in program evaluation is limited. In this study, we used LTA to evaluate a school-based mental health promotion and violence prevention program for children. Multiple mental health outcome measures (emotional symptoms, conduct problems, hyperactivity/inattention, peer relationship problems, and lack of prosocial behaviour) were collected before and after intervention. The nested nature of data (students nested in schools) required that we take into consideration the dependence of children within classrooms. We evaluated program effects by comparing transition patterns of participants in the intervention and control groups. We derived profiles based on multiple mental health outcome measures, and then used the difference across treatment groups in the probability of transitioning between risk profiles to estimate a program effect. Below we describe the program evaluation design and outcome data. Then, we present evaluation results using LTA. We conclude by discussing the strengths and limitations of LTA.

1 Method

1.1 Background of PAX Program

In February 2012, the Manitoba Government launched the first province-wide pilot of PAX Good Behavior Game (PAX GBG or PAX for short) offering Grade 1 teachers the training and tools to help children develop social, emotional and self-regulation skills. PAX is a classroom strategy that improves children's self-regulation and ability to delay rewards in multiple peer contexts, during school activities, thereby creating a safer environment that is conducive to learning, positive peer interactions, and immediate and distal education outcomes and self-regulation in children (Barrish, Saunders & Wolfe 1969; Bradshaw, Zmuda et al. 2009; Embry 2002, 2011; Ialongo et al. 1999). Besides decreasing disruptive and hyperactive behaviour, the intervention can also decrease children's anxiety or emotional arousal about school (Flannery et al. 2003).

1.2 Participants and Design

The provincial pilot study included about 200 schools from nearly every school division in Manitoba, Canada, including First Nation and independent schools. Schools were randomly assigned within school divisions to implement PAX in either 2011/12 (treatment schools) or the following school year (waitlist control schools). Figure 1 presents the selection and flow of clusters and individual participants through the randomized trial. The pilot involved about 5000 students and their teachers. Schools were asked to facilitate the collection of pre- and post-program outcome measures of child mental health at the beginning and end of the school year.

1.3 Measurement

Pre-test and post-test data were collected from classroom teachers using the Strengths and Difficulties Questionnaire (SDQ), one of the most widely used and well-validated measures of children's mental health (Goodman 2001). SDQ measures Emotional Symptoms (5 items: e.g., feeling anxious or depressed), Conduct Problems (5 items: e.g., bullying other children), Hyperactivity/Inattention (5 items: e.g., restless, easily distracted), Peer Relationship Problems (5 items: e.g., having few friends, being bullied by other children), and Prosocial Behaviour (5 items: e.g., sharing with and helping others). Each item was rated by teachers using a 3-point Likert scale: not true, somewhat true, or certainly true. Internal consistency for all five SDQ scales was acceptably high within the current sample across time points (Cronbach's alphas: mean = .83; range = .72–.90).

2 Results

As with most longitudinal studies and illustrated in Fig. 1, the problem of missing data occurred in our study. Though we had roughly equal numbers of schools between PAX and control groups, the control group was less likely to complete data assessments than the PAX group. However, a series of Wilcoxon and t-tests indicated that the sample with the completed data assessments was representative of the original sample at all time points in terms of teachers' ratings of mental health

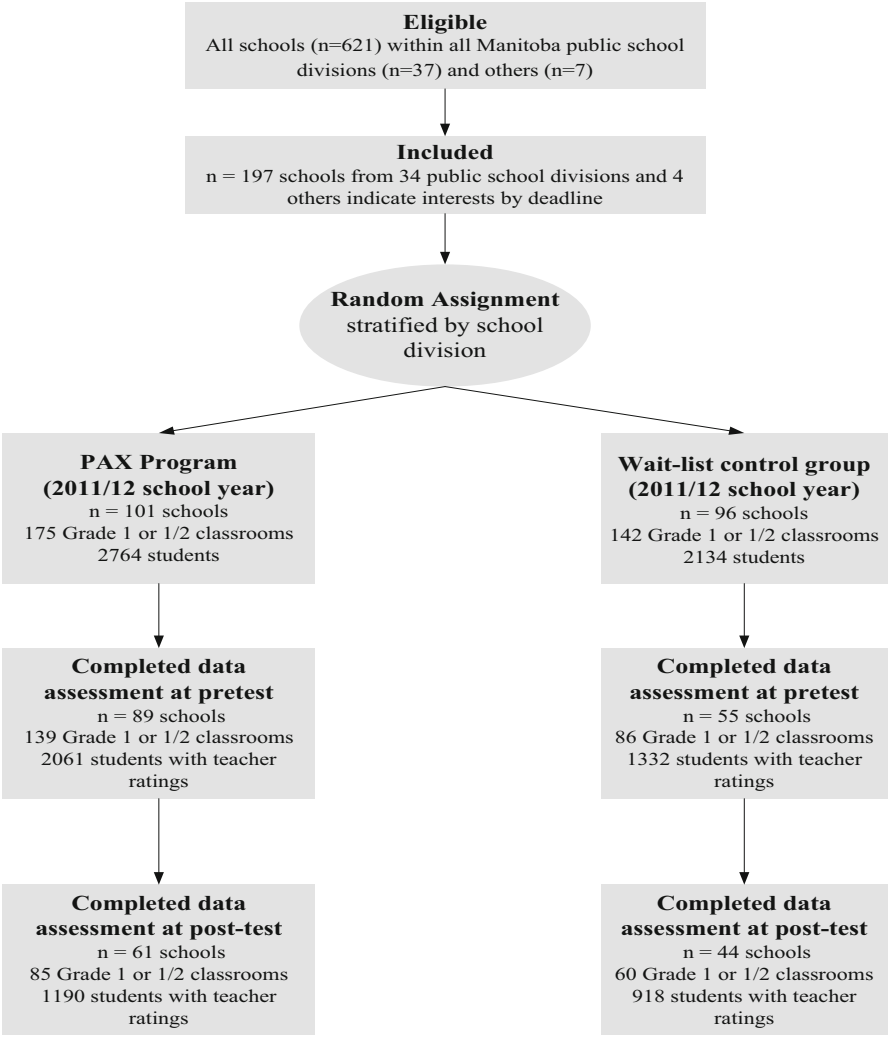


Fig. 1 Selection and flow of clusters and individual participants through the randomized trial

Table 1 Descriptive statistics

	Intervention (N = 2061)	Control (N = 1332)
	M(SD)	M(SD)
Age	7.04(0.37)	7.02(0.33)
	n (%)	n (%)
Gender		
Male	811 (49.6)	626 (49.6)
Female	824 (50.4)	637 (50.4)
Aboriginal status		
Yes	682 (40.8)	341 (30.5)
No	989 (59.2)	779 (69.5)
English as first language		
Yes	1689 (89.5 %)	1410 (85.4 %)
No	199 (10.5 %)	241 (14.6 %)
	M(SD)	M(SD)
T1 Prosocial behavior	6.99(2.70)	7.27(2.56)
T2 Prosocial behavior	7.81(2.58)	7.55(2.41)
T1 Emotional symptoms	2.06(2.29)	1.47(2.07)
T2 Emotional symptoms	1.41(1.91)	1.39(1.99)
T1 Conduct problems	1.54(2.23)	1.28(1.99)
T2 Conduct problems	1.18(1.20)	1.17(1.91)
T1 Hyperactivity	4.06(3.33)	3.59(3.25)
T2 Hyperactivity	3.25(3.26)	3.23(3.21)
T1 Peer relationship problem	1.86(2.06)	1.46(1.89)
T2 Peer relationship problem	1.35(1.86)	1.30(1.85)

Note: T1 refers to data collected at pre-test and T2 refers to data collected at post-test

and demographics. To ensure that LCA and LTA analyses were based on at least one data point, we selected only those students with completed SDQ at pretest. The sample available for LTA model included 3393 participants from 158 schools. Descriptive statistics are outlined in Table 1.

LCA was used first to define latent class structures for both intervention and control groups and separately for both pre-test (Time 1) and post-test (Time 2) assessments. Then the LTA model was fit to the combined intervention and control groups to examine the probability of participants transitioning among the defined latent classes between two data collection waves. These transitions are conditional on treatment assignment. Because the intervention was implemented in classroom-based groups in first grade or mix of first and second grades, classroom-based dependency in scores might affect the outcome of conventional LTA and LCA analyses. Therefore, multilevel LTA and LCA will be conducted. Because the randomization of treatment conditions was at school level and most schools have one or two classrooms, the unit of the multilevel analyses is school. There are several approaches to the analysis of multilevel data (Muthén & Satorra 1995).

Our approach for the multilevel LTA and LCA is to compute standard errors taking into account non-independence of observations due to cluster sampling (students nested in schools).

2.1 Cross-Sectional Analysis with LCA

LCA models with the five SDQ scales as indicators were estimated to determine whether and how many meaningful mental health risk classes of individuals were in each group (intervention vs. control) and each time point (pre-test and post-test), and if patterns in response probability and class proportions were similar across groups and consistent with theory.

Determining the number of latent classes within LCA remains one of the challenging issues. Currently, determining the number of classes consists of finding the model with the smallest Bayesian information criterion (BIC) (Kass & Raftery 1995) and a significant likelihood ratio test (LMR-LRT) (Lo, Mendall & Rubin 2001). LMR-LRT provides a statistical comparison of the fit of a given model with a model with one less class. In addition to considering these indices of fit, the optimal number of classes should be determined by a combination of factors that includes the research question, parsimony, theoretical justification, and interpretability (Bauer & Curran 2003). As in most LCAs, each model was tested in an iterative fashion. A one-class solution was fit first and then one class was added at a time until the model fit the data well. One- to 4-class solutions were tested on the data. Table 2 lists the BIC values for the 1-class to 4-class solutions for each of the four LCA models. Application of the maximum BIC for model selection in our case did not result in an overwhelmingly clear determination of the "best" model and the BIC continued to improve as more groups were added. The LMR-LRT test of model fit indicated that the increment of estimate from a 3-class model to a 4-class model was not significant. Thus, the 3-class model was chosen as optimal in that it best balanced goodness-of-fit and parsimony. Table 2 suggests a 3-class profile model fit the data for both control and intervention groups and at both pre-test and post-test.

Table 2 BIC values and LMR-LRT test for different numbers of classes

	# of classes	BIC	LMP-LRT	BIC	LMP-LRT
Pre-test		Intervention (N = 2061)		Control (N = 1332)	
	1	48024.6	–	30086.6	–
	2	44832.1	<.001	28046.0	<.001
	3	43811.9	.005	27436.6	.02
	4	43292.6	.22	27099.4	.09
Post-test		Intervention (N = 1190)		Control (N = 917)	
	1	26599.7	–	20410.4	–
	2	24474.1	<.001	18975.3	<.001
	3	24001.1	0.01	18555.1	0.05
	4	23667.3	0.08	18348.7	0.33

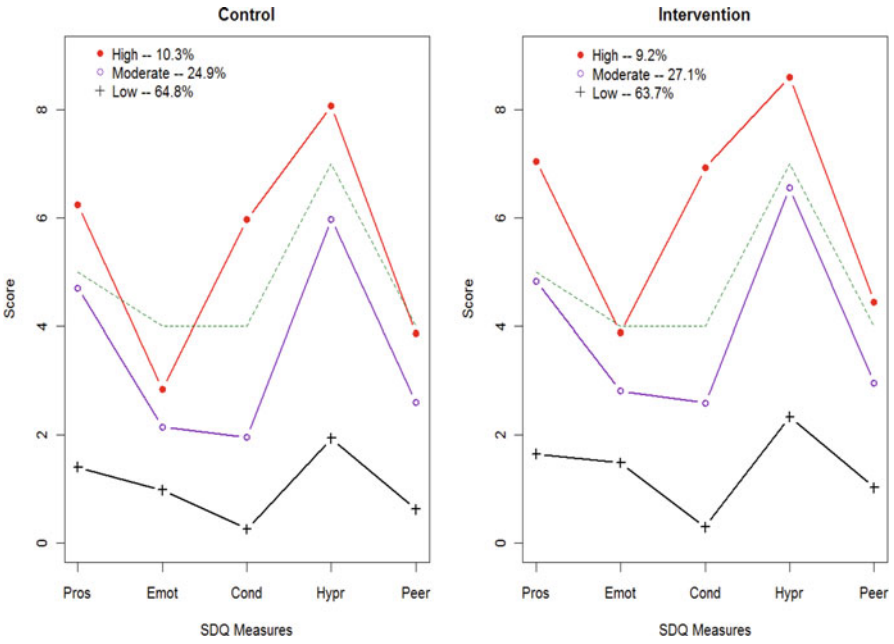


Fig. 2 Response profile at pre-test. *Note:* (1) Prosocial scale was reversed in LCA models with higher level indicating lack of pro-social skills. (2) *Pros* lack of Prosocial skills, *Emot* emotional symptoms, *Cond* conduct disorder, *Hypr* hyperactivity, *Peer* peer problems. (3) The *dotted line* represents the cut-offs for abnormal antisocial and prosocial behaviors suggested by Niclasen et al. 2012

Children were assigned to a latent class on the basis of their highest estimated class probability. Symptom endorsement profiles for the 3-class models at pre-test are presented in Fig. 2. The endorsement profiles of children were highly comparable across the three classes. Children in Class 1 exhibited a high mean level of antisocial behaviours and low mean level of prosocial behaviours. Class 1, accounting for 10.3 % of the intervention group and 9.2 % of the control group, was called the high-risk class. Figure 2 shows that children in the high-risk group have the greatest likelihood of abnormal antisocial behaviours or pro-social skills according the cut-offs suggested by Niclasen et al. (2012). Children in Class 2 exhibited a modest level of antisocial behaviours and prosocial skills. Class 2, accounting for 24.9 % of the intervention group and 27.1 % of the control group, was called the *moderate-risk class*. The largest group, called the *low-risk class*, comprised children who rarely exhibited any antisocial behaviours and showed good prosocial skills. This group is estimated to account for 64.8 % of the intervention group and 63.7 % of the control group.

The risk profiles of children were quite similar across the intervention and control groups before the PAX intervention program. To further examine the similarity between intervention and control groups before intervention, we also extended the

Table 3 Results for predictors of latent class membership

	Intervention (N = 2061)	Control (N = 1332)
Variable	Odds Ratios (95 % CI)	Odds Ratios (95 % CI)
Moderate risk		
Female vs. male	0.45(0.34–0.60)***	0.48(0.36–0.65)***
Aboriginal vs. non-aboriginal	2.91(2.07–4.08)***	3.71(2.08–6.63)***
High risk		
Female vs. male	0.42(0.29–0.62)***	0.42(0.25–0.71)***
Aboriginal vs. non-aboriginal	2.19(1.16–4.16)*	3.48(2.10–5.77)***
Low risk		

Note: The reference category is the low risk group. *p < .05; **p < .01; ***p < .001

LCA models by including predictors of class membership (gender and aboriginal status). Results are shown in Table 3. At pre-test, male students have greater odds of being in the high and moderate mental health classes than female students. Aboriginal participants have greater odds of being in the high and moderate mental health classes than non-aboriginal participants. These predictions are quite consistent for both intervention and control groups.

2.2 Longitudinal Analysis with LTA

The 3-class LCA model was extended to LTA for the combined control and intervention samples. The LTA framework is shown in Fig. 3. In this model, the risk profiles were constrained as equal across time to assure measurement invariance (i.e., the class structure stays the same across time). This measurement invariance assumption makes it possible to interpret the group difference in the transition patterns among the defined latent classes from pre- to post-test collection waves as intervention effectiveness. Transition probabilities estimated from LTA are reported in Table 4 and provide information about an individual's latent class status at post-test given their latent status at pre-test. Table 4 results show change over time in class membership for some participants, mostly for those in the intervention group. Most participants in the control group stayed in the same class from pre-test to post-test. In the control group, participants in the low risk class had a 98.9 % probability of remaining in the low risk class at post-test, while 85 % of students who were in the moderate-risk class at pre-test remained in the moderate-risk class at post-test, and 74.4 % of students in the high-risk class at pre-test remained in the high-risk class at post-test. For the control group, only 7 % of students would move to a different latent risk class from pre-test to post-test.

Wald chi-square tests were performed to compare the transition probabilities across the intervention and control groups. Transition probabilities were somewhat different for participants in the intervention group: almost all students (97.3 %) in the low-risk class remained in the low-risk class from pre-test to post-test,

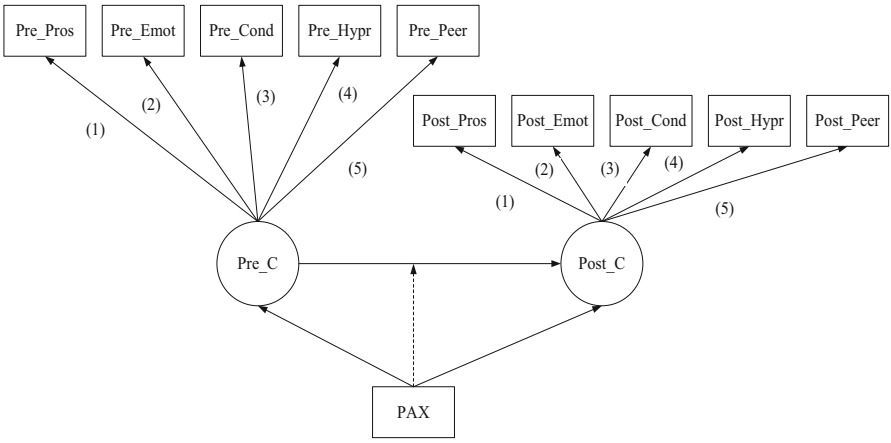


Fig. 3 Framework of Latent Transition Analysis (LTA) Model. *Note*: (1) Prosocial scale was reversed in LTA models with higher level indicating lack of pro-social skills. (2) *Pros* lack of Prosocial skills, *Emot* emotional symptoms, *Cond* conduct disorder, *Hypr* hyperactivity, *Peer* peer problems; Pre: Pretest; Post: Post test. (3) In this model, the means of the latent class indicators for a given class are held equal for the two categorical latent variables across two times. The (1–5) use the list function to assign equality labels to these model parameters

Table 4 Transition probabilities from pre-test to post-test

	Post-test					
	Intervention (N = 2061)			Control (N = 1332)		
Pre-test	Low	Moderate	High	Low	Moderate	High
Low	.973	.025	.002	.989	.011	0
Moderate	.351	.576	.073	.092	.850	.058
High	.086	.361	.553	0	.256	.744

Note: Cell entries are the predict probabilities of latent class membership at post-test given their latent status at pre-test
For example, for the intervention group, 35.1 % of participants who were in the moderate risk class at pre-test were predicted to transition to the low risk class at post-test

and moderate-risk class students had only a 57.6 % probability of remaining in the moderate-risk class from pre-test to post-test and a 35.1 % probability of transitioning from moderate-risk to low-risk from pre-test to post-test. In the intervention group, students had only a 55.3 % probability of remaining in the high-risk class from pre- to post-test and a 44.7 % probability of moving favourably from the high-risk class to the low/moderate-risk from pre-test to post-test.

The net effect of the PAX program for moderate-risk children improving over time ($0.351 - 0.092 = 0.259$) was significant ($p < .001$). Moderate-risk children in the PAX group were nearly 5.3 times more likely to improve over time than those in the control group. The net effect of the PAX program for the high-risk children

improving over time ($0.744 - 0.553 = 0.191$) was also significant ($p < .001$). The high-risk children in the PAX group were nearly 3.8 times more likely to improve over time than those in the control group.

3 Discussion

Implementing and evaluating an intervention program on a large scale under real world conditions requires prolonged and intensive efforts, and often the overall observable effects are modest, at best. Also in school-based mental health intervention studies, one size does not fit all—there are many forms of unobserved heterogeneity among participants and only some subgroups within the overall sample may demonstrate observable effects. These subgroup intervention effects might be obscured with the traditional *variable-oriented* statistical approaches. In this paper, we illustrated how *person-oriented* statistical approaches such as latent transition analysis (LTA) can help us reveal subgroup intervention effects.

LTA has many advantages for program evaluation. In LTA, multivariate normal data are not required. This makes it useful in studies where outcome variables are continuous measures with a large number of observed values clustered at zero. Indicators for LTA do not have to be at a continuous level other than nominal (Collins & Lanza 2010). LTA can also provide information about participant groups that benefited from an intervention even if the sample as a whole did not appear to benefit. In order to identify which subgroups of participants might benefit differentially from the intervention, one can form risk status categories using cut-offs on the pre-intervention risk scores, or use LTA or LCA. The former approach has some advantages, in that the subgroups are guaranteed to be meaningful if they are based on theoretical and empirical grounds. However, when there are multiple outcomes, it might be very challenging or impossible to use. Advantages of LTA include identifying different response patterns in which participants are classified to (latent) classes directly by the model. For multiple pre-intervention measures, van Lier, Muthen, van der Sar, and Crijnen (2004) have shown that the use of LCA improves predictive accuracy of risk status. Applying these two different approaches can dramatically impact the effect size estimates and future intervention design. In addition, LTA allows for measurement error so that individuals who do not map directly into a class are dealt with in a systematic way.

Similar to other statistical analyses, LTA has limitations. It requires large sample sizes: when sample size is small, where one of the latent classes has a very low prevalence, or when membership in one of the classes is essentially zero for some level of a covariate, the estimates (especially standard errors) are not reliable or sometimes cannot be estimated. LTA models are also subject to misspecification and unobserved heterogeneity. We recommend trying a variety of scenarios and planning primary analyses in advance. If an entire sample shows a homogeneous underlying pattern of change, with some variation around the single pattern, then a conventional statistical approach (e.g., regression analysis) is preferable. If, on the other hand,

there is a moderate intervention effect overall, and the effect varies as a function of the initial risk level, then using a *person-oriented* approach is recommended. It is well-suited to answering questions for understanding outcome change for subgroups across discrete qualitative states.

References

Barrish, H. H., Saunders, M., & Wolfe, M. D. (1969). Good behavior game: Effects of individual contingencies for group consequences on disruptive behavior in a classroom. *Journal of Applied Behavior Analysis, 2*(2), 119–124.

Bauer, D., & Curran, P. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods, 8*, 338–363.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods, 1*, 3–29.

Bradshaw, C. P., Zmuda, J. H., et al. (2009). Longitudinal impact of two universal preventive interventions in first grade on educational outcomes in high school. *Journal of Educational Psychology, 101*(4), 926–937.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis with applications in the social, behavioral, and health sciences*. Hoboken, NJ: Wiley.

Collins, L. M., & Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research, 27*(1), 131–157.

Embry, D. D. (2002). The good behavior game: A best practice candidate as a universal behavioral vaccine. *Clinical Child & Family Psychology Review, 5*(4), 273–297.

Embry, D. D. (2011). Behavioral vaccines and evidence-based kernels: Nonpharmaceutical approaches for the prevention of mental, emotional, and behavioral disorders. *Psychiatric Clinics of North America, 34*, 1–34.

Flannery, D. J., Vazsonyi, A., Liau, A., Guo, S., Powell, K., Atha, H., et al. (2003). Initial behavior outcomes for Peacebuilders universal school-based violence prevention program. *Developmental Psychology, 39*, 292–308.

Goodman, R. (2001). Psychometric properties of the Strengths and Difficulties Questionnaire (SDQ). *Journal of the American Academy of Child and Adolescent Psychiatry, 40*, 1337–1345.

Ialongo, N., Werthamer, L., Kellam, S. G., Brown, C. H., Wang, S., & Lin, Y. (1999). Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression, and antisocial behavior. *American Journal of Community Psychology, 27*(5), 599–641.

Jiang, D., Pepler, D., & Yao, H. (2010). The effect of population heterogeneity on statistical power in the design and evaluation of interventions. *International Journal of Behavioral Development, 34*(5), 473–480.

Kass, R. E., & Raftery, A. E. (1995). Bayes factor. *Journal of the American Statistical Association, 90*, 773–795.

Lo, Y., Mendall, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*, 767–778.

Muthén, B., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. Marsden (Ed.), *Sociological methodology* (pp. 216–316). Boston: Blackwell.

Niclasen, J., Teasdale, T. W., Andersen, A. M., Skovgaard, A. M., Elberling, H., & Obel, C. (2012). Psychometric properties of the Danish strength and difficulties questionnaire: The SDQ assessed for more than 70,000 raters in four different cohorts. *PLoS One, 7*, e32025.

Pastor, D. A., Barron, K. E., Miller, B. J., & Davis, S. L. (2007). A latent profile analysis of college students' achievement goal orientation. *Contemporary Educational Psychology, 32*(1), 8–47.

Thompson, A. M., Mary, R. J., & Fraser, M. W. (2011). Assessing person-centered outcomes in practice research: A latent transition profile framework. *Journal of Community Psychology, 39*(8), 987–1002.

van Lier, P. A. C., Muthen, B. O., van der Sar, R. M., & Crijnen, A. (2004). Preventing disruptive behavior in elementary schoolchildren: Impact of a universal classroom-based intervention. *Journal of Consulting and Clinical Psychology, 72*(3), 467–478.

The Theory and Practice of Personality Development Measurements

Wei-Dong Wang, Fan Feng, Xue-Yu Lv, Jin-Hua Zhang, Lan Hong, Gui-Xia Li, and Jian Wang

Abstract To determine the structure of memory-tracing developmental levels, we created the Wang Wei-dong Memory-Tracing Personality Development Inventory (WMPI) based on the perspective of abnormal personality development theory in Chinese medical psychology. We used literature analysis, qualitative research, and our own analysis to build the theoretical basis and structure of the WMPI and compiled items while considering traditional Chinese medicine and psychology. We also assessed the reliability and validity of the inventory by means of explorative factor analysis and confirmatory factor analysis. The final WMPI was comprised of 9 subscales, 37 dimensions, and 248 items and it was divided into childhood, adolescence, and adulthood stages. The reliability of the entire inventory was 0.990, and the reliability of the 9 subscales was between 0.780 and 0.963. The RMSEA of every subscale was less than 1, and the NNFI and CFI were nearly 0.90, which indicated the inventory had good quality. The reliability and validity tests demonstrated holism and the developmental viewpoint of traditional Chinese medicine, which played a guiding role in the process of compiling the WMPI. The WMPI has good reliability and construct validity.

Keywords Personality development • Memory-tracing • WMPI

1 Background and Purpose

According to the theory of personality, scholars have made great efforts to describe and clarify personality structure, such as the 16 factors model (Catter, 1979) and the big five (McCrae & Costa, 1999). However, they are still of limited value in clinical practice (Murray, 1938; Pervin, 1990; Ruston & Irwing, 2008) because psychological measurement mainly focuses on the present psychological condition of the subjects and not on the process of psychological development.

W.-D. Wang • F. Feng • X.-Y. Lv • J.-H. Zhang • L. Hong • G.-X. Li • J. Wang (✉)
Guang'anmen Hospital, China Academy of Chinese Medical Sciences, No. 5 North Line Xicheng District, 100053 Beijing, China
e-mail: wjmd@263.net

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_29

1.1 The Theory of Psychological Development in Traditional Chinese Medicine (TCM)

TCM theories (Wang & Yang, 2013) suggest that all substances, including nature, are eternally and endlessly in movement. To "move" is the fundamental law of nature and the various phenomena of nature, such as life, health, and disease, are all forms of material movement. TCM psychologists have indicated that personality is the overall dynamic and dialectical development of the system. Personality continues from an individual's birth to death, however, at the same time, it also has phases. Time, phase, and the elements of psychological development is said to form the interchange, which is a type of internal reticular structure.

TCM holds that personality is a dynamically balanced system, and a personality follows the principle of equilibrium between yin and yang. TCM views the development of an individual's personality and characteristics from the perspective of dynamic development. It states personality in a lifetime has different features at different stages, and its rich elements present a spindle structure. The degrees of development of various elements in the era of youth and middle age are the most plentiful, that is childhood personality development degrees increase, and in the elderly, the elements of personality development gradually weaken. These theories provide the reference for TCM psychology to explore personality development.

1.2 The Theory of Abnormal Personality Development in TCM

Based on the views above, Professor WangWei-dong described abnormal personality development (Wang, Du, Lv & Li 2012). It was assumed that the personality is composed of personality elements. The personality elements are related to each other and are influenced by the natural and social environment. During a lifetime, the personality and personality elements keep changing, however, they maintain special characteristics during particular periods. An abnormal personality is the result of abnormal development. Other theoretical sources have also provided contributions to the understanding of abnormal personality development, such as: (1) psychodynamic theory; (2) developmental psychology theory; (3) clinical experience and theory; and (4) "regrowth treatment" (Wang, 2012b), the effectiveness of clinical practice.

Personality structure within abnormal developmental psychology theory is based on the understanding of personality formation as follows: (1) Personality is the result of dynamic changes and the main body gradually stabilizes in the process of the person's psychological development factors and interactions; (2) although the formation of personality is a process, to a certain extent it has periodicity; and (3) the interaction processes of psychological elements, the mental development dynamic changes in internal structure, and the body's gradual stabilization are the inner motives underlying personality formation. The behavioral characteristics that can be

observed are external performances. While at a certain age, personality is relatively stable, there are subtle changes. In terms of an individual's life, personality is a process of constant development.

1.3 Causes and Types of Abnormal Psychological Development

The formation of mental diseases is not simply about the state of the disease itself. Normal psychological development should be understood as the result of an individual's psychological development process. Abnormal psychological development and the formation process of psychological diseases must include normal development with deviation and absence (Wang, 2012a, 2012b).

1. Psychological diseases caused by abnormal development with deviation are processes in which normal psychological development appears to have migrated and the result was disease.
2. Abnormal development with absence includes the absence of normal mental development, including both the time dimension and space dimension. Normal psychological development includes the complete development of each factor in the spatial dimension and sustainable development in the time dimension. Growth absence mainly manifests in two aspects; that is, the lack of "growth factors" and missing "growth stages."

On the basis of abnormal personality development theory, we constructed the Wang Wei-dong Memory-Tracing Personality Inventory (WMPI) to measure personality elements at different ages to describe the process of personality development and formation mechanisms of abnormal personalities.

2 Methods

2.1 Homework Analysis

To collect information for psychotherapy during clinical sessions, a type of homework, which was called outline homework, was given to the patients.

The details of the homework are as follows:

Please try to remember the important events of your lifetime and write them down according to the outline. This exercise is very important for your practician to understand your condition and to help you get better. It has to be done alone and cannot be shown to anyone else. Your homework will remain anonymous and secret.

1. *The painful or sad events and injustices you experienced.*
2. *The horrible, frightening events and worries you experienced.*
3. *Sex or affection related events you find hard to talk about.*
4. *The person(s) who has/have been the most trustful, reliable and unforgettable and why?*
5. *The most relaxed, joyful and happy period(s) of your life and why?*

During clinical practice for approximately 20 years, we collected homework from almost 300 patients with mental disorders, and we selected 150 pieces of complete homework for qualitative analysis. In combination with clinical experience, we identified the following nine personality factors: (1) life events; (2) parenting styles; (3) way of thinking; (4) courage; (5) ego consciousness; (6) interpersonal relationship; (7) volition; (8) sex development; and (9) world conception. These nine factors are the subscales of the WMPI, and at the same time they are the basic personality elements. The life events and parenting styles are exterior elements and all other factors are interior elements. Both exterior and interior elements affect the route of personality development.

2.2 Literature Analysis

According to the literature analysis, the ability to distinguish real and surface emotions develop rapidly at 4 years old, and become stable after 5 years old (Zhang, 2011), and independence develops fast between 3 years to 5 years old (Hei, 2008). The results of a study on the stability of personality indicated that personality differences appear after 5 years old (Gao & Yang, 2007). Research on creative personalities has indicated that 4 years old is an important period for the development of creativity (Lu, 2007). Another research study on Chinese children showed personality, including prosocial ability, intelligence, extroversion and emotional stability, grows rapidly at 6–7 years old (Zhuo, 2008). Summing up the above, and considering our clinical experience, we regarded 3–7 years old as the key period of personality development.

Conway and Pleydell-Pearce (2000) suggested that autobiographical memory could be represented at three levels: lifetime periods, general events, and event-specific knowledge. Event-specific knowledge is important because it includes real feelings about events and has great influence on people. The patients' reports concerning their life events, experiences, and feelings are rather similar to autobiographical memories. Therefore, memory-tracing research (Wang et al., 2012) contains autobiographical memory, especially event-specific knowledge. The memory extraction style of WMPI is similar to autobiographical memory (Conway & Pleydell-Pearce, 2000; Matuszewski et al., 2006). The WMPI focuses on the events and feelings in subjects' autobiographical memory instead of the real situation because autobiographical memory affects the personality more deeply.

2.3 Item Compiling

Some items were constructed after discussion among group members, some items were selected from the homework of the patients through qualitative analysis, and some items came from other similar inventories or scales. We used a 5-point Likert scale to evaluate the occurring frequencies using positive and negative scoring.

WMPI aims to measure the personality element development levels at 3 important age periods: 3–6 years old; 7–18 years old and 19–25 years old. At the age of 26, it was assumed that the personality is stabilized. We asked the subjects to evaluate their life events, cognitions, and behaviors in these three periods.

2.4 Item Selection

We solicited opinions from psychological experts and patients with mental disorders, people from other fields and people with different educational levels, and we repeatedly modified the items based on these results. Experts considered clinical data as first-hand information, and they thought it could directly reflect the patient's experience and process of psychological development. We therefore designed the questionnaire subscales based on the clinical data. According to the results of operation outline qualitative analysis, combining modern personality theories and the psychological characteristics of different age stages, experts thought it was rational that the questionnaire was divided into 9 subscales, 41 dimensions, and three age periods. They also suggested that some ambiguous items needed to be revised. At the same time, we collected opinions from normal people and outpatients in a psychological clinic. According to their feedback, we deleted the items that were hard to understand and were considered meaningless. Thus, we obtained an original version of the WMPI consisting of 352 items, 41 dimensions, and 9 subscales.

2.5 Testing

We carried on two tests of the WMPI. In the first test, we used the original version of the WMPI questionnaire and performed an item analysis and explorative factor analysis on these test data. Through modifications of subscales, dimensions, and items on the basis of the analysis, we created the formal questionnaire. We then carried out the second tests to obtain the formal version, and performed confirmatory factor analysis to verify the rationality of the scale structure.

2.6 Participants

The participants in the first test were healthy controls and diagnosed mental patients. The WMPI was converted into software and was uploaded to the internet. Through a convenience sampling method, the inventory data were collected through the network. A total of 5611 people took part in the testing, and 1474 completed it. After screening for false responses, 1151 pieces of effective inventory remained. Among

the effective inventory, there were 984 pieces from normal people and 167 pieces from mental patients. Participants in the second test involved 198 health controls and 95 mental patients after screening.

2.7 Statistical Analysis

We conducted item analysis, explorative factor analysis, confirmatory factor analysis, and reliability analysis. Item analysis and exploratory factor analysis were conducted with SPSS20.0. AMOS19.0 was used to analyze the construct validity.

3 Results

3.1 Discriminability Analysis

To construct a scale that discriminated between normal and abnormal personality development, we tested whether the mean item score for each item was the same for patients and healthy controls using t tests for independent samples. We deleted the 27 items that were significant using a nominal Type I error rate of 0.05.

3.2 Exploratory Factor Analysis

The preliminary WMPI contained 3 periods and 9 subscales, and each subscale had several dimensions. Because of the complicated framework and large amount of items, we conducted exploratory factor analysis on each subscale rather than on the whole inventory. Through Eigenvalue analysis and fixed factors analysis, we deleted the 81 items with factor loadings less than 0.3 and then modified the dimensions.

3.3 Formal Edition of the WMPI

Using the results of item analysis and explorative factor analysis, we discussed, modified and deleted some items. The formal edition of the WMPI contained 248 items clustered into 37 dimensions, which were clustered into nine subscales (Table 1).

Table 1 The construction of WMPI

Subscales	Dimensions	Amount of items
Courage	Interpersonal fear	12
	Natural fear	4
	Adaptability	7
	Anxiety	5
Ego consciousness	Social ego	6
	Physiological ego	4
	Family ego	3
	Independence	6
	Self-care capability	4
Way of thinking	Abnormal thoughts	10
	Irrational thoughts	5
	Caution	3
	Hybris	5
Volition	Resolution	4
	Consciousness	3
	Delay of gratification	4
	Insistence	4
Interpersonal relationship	Gregariousness	10
	Altruism	4
	Dependence	5
Sex development	Relationship with opposite sex	8
	Cognition of love	10
	Cognition of sex	4
World conception	Motivation and attribution	9
	Values	11
	Viewpoint of cause	4
	Viewpoint of friendship	4
	Viewpoint of health	5
Life events	Family events	7
	Social events	19
	School events	10
	Events relate to sex	9
Parenting styles	Stern punishment	9
	Excessive interference	10
	Spoiling	3
	Contradictory parenting	5
	Ignore parenting	3
Lie detection	–	10
Whole inventory	–	248

Table 2 The values of Cronbach's alpha for each subscale and for each age period

Item	3–6 years old	7–18 years old	19–25years old
Courage	.878	.879	.877
Ego consciousness	.903	.905	.909
Way of thinking	.894	.884	.871
Interpersonal relationship	.780	.821	.831
Volition	.867	.894	.891
Life events	.955	.962	.963
Parenting styles	.877	.873	.880
Sex development	–	.830	.963
World conception	–	.868	.880

3.4 Reliability Analysis

Cronbach's coefficient alpha for the whole inventory was equal to 0.990. Table 2 shows the values of Cronbach's alpha for each subscale and for each age period.

3.5 Construct Validity Analysis

The confirmatory factor models used to investigate the construct validity showed relatively good fit. Ego consciousness for 3–6 years old, sex development for 7–18 years old and 19–25 years old were relatively lower. However, after being considered comprehensively, we decided to continue using them. The results of construct validity analysis are shown in Table 3.

4 Discussion and Conclusion

The WMPI is based on a holistic view representative of Chinese philosophy and TCM. From a holistic view, humans and the environment exist as a whole. During the process of constructing the WMPI, we focused on the relationship between humans and the environment. We considered life events and parenting styles as exterior elements affecting personal development. It is conjectured that these exterior elements (partly) cause abnormal personalities and mental disorders.

The WMPI was used to collect information on subjects who were less than 25 years old. This information can help clinical psychologists to understand the growing experience in a short time. Researchers can discover the processes of personality development and infer the mechanisms of mental disorders.

Table 3 The results of construct validity analysis

Ages	Dimension	GFI	AGFI	CFI	RMSEA
3–6 years old	Courage	.863	.832	.763	.091
	Ego consciousness	.837	.794	.735	.105
	Way of thinking	.920	.917	.907	.049
	Volition	.980	.968	.975	.044
	Interpersonal relationship	.913	.872	.876	.092
	Life events	.859	.838	.875	.066
	Parenting styles	.895	.868	.868	.069
7–18 years old	Courage	.904	.885	.883	.056
	Ego consciousness	.842	.798	.784	.097
	Way of thinking	.961	.985	.958	.047
	Volition	.971	.955	.968	.050
	Interpersonal relationship	.901	.860	.867	.091
	Sex development	.751	.670	.669	.128
	World conception.	.823	.775	.766	.089
	Life events	.731	.705	.756	.073
	Parenting styles	.801	.767	.781	.082
19–25 years old	Courage	.896	.876	.874	.058
	Ego consciousness	.832	.785	.784	.098
	Way of thinking	.945	.982	.913	.045
	Volition	.975	.962	.973	.045
	Interpersonal relationship	.907	.867	.866	.087
	Sex development	.760	.682	.639	.126
	World conception.	.821	.769	.771	.090
	Life events	.763	.740	.795	.071
	Parenting styles	.800	.764	.792	.085

We report here criterion validity for the WMPI. Because we obtained scores with the SCL-90 Chinese version (Wang, 1984), these scores from the SCL-90 could have been used as criteria. Although the WMPI and SCL-90 scores were highly correlated, the SCL-90 still cannot be used as a criterion for the WMPI because the theoretical foundation of both instruments is different. The SCL-90 was designed as a symptom scale to differentiate between mental patients and healthy controls.

Through repeated discussions and trials, the WMPI was developed to be an inventory that has good reliability, construct validity, and congruent validity.

References

Catter, R. B. (1979). *Personality and learning theory*. New York, NY: Springer.

Conway, M. A., & Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological Review, 107*, 261–288.

Gao, W., & Yang, L. Z. (2007). 3–9 sui er tong renge te zhi wen ding xing li jie de fa zhan te dian [Developmental features of understandings about the stability of traits among Chinese children aged from 3 to 9]. *Psychological Development and Education, 3*, 6–12.

Hei, L. J. (2008). 3–5 sui er tong zizhuxingjiegouji fa zhan te dianyanjiu [Study on independence construction and development characteristics among Chinese children aged from 3 to 5]. Unpublished report, Liaoning Normal University, Dalian, China.

Lu, Q. (2007). 3–5 sui you erzhuangzaoxingrenge de jiegou fa zhan te dianyu lei xing [Construction, developmental features and type of creative personality among Chinese children aged from 3 to 5]. Unpublished report, Liaoning Normal University, Dalian, China.

Matuszewski, V., Piolino, P., de la Sayette, V., Lalevée, C., Pélerin, A., Dupuy, B., et al. (2006). Retrieval mechanisms for autobiographical memories: Insights from the frontal variant of frontotemporal denaentia. *Neuropsychologia, 44*, 2386–2397.

McCrae, R. R., & Costa, P. T., Jr. (1999). A five-factor theory of personality. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 139–153). New York, NY: Guilford Press.

Murray, H. A. (1938). *Explorations in personality*. Oxford, UK: Oxford University Press.

Pervin, L. A. (1990). *Handbook of personality: Theory and research*. New York, NY: Guilford Press.

Ruston, J. P., & Irwing, P. (2008). A general factor of personality from two meta-analyses of the big five. *Personality and Individual Differences, 45*, 679–683.

Wang, Z. Y. (1984). Zheng zhuang zi ping liang biao [Symptom Checklist 90]. *Shanghai Archives of Psychiatry, 2*(69–70), 93–95.

Wang, W. D. (2012a). *Low resistance thought induction psychotherapy: A guide to theory and practice*. Shelton, CT: People's Medical Publishing House (PMPH).

Wang, W. D. (2012b). *Developmental psychotherapeutics: A theoretical system of psychotherapy based on abnormal development*. Shelton, CT: People's Medical Publishing House (PMPH).

Wang, W. D., Du, H., Lv, X. Y., & Li, S. T. (2012). Discussion about clinical recallable and retrospective study of psychic and psychological illness. *Medicine and Philosophy, 33*, 22–23.

Wang, K. Q., Yang, Q. L. (2013). *Zhong yi xin li xue ji chu li lun* [The basic theory of psychology in traditional Chinese medicine]. Shelton, CT: People's Medical Publishing House (PMPH).

Zhang, J. R. (2011). 3–12 sui er tong ren ge de jie gou ping ding ji qi fa zhan te dian de zhui zong yan jiu [The follow-up study on constructive evaluation and development characteristics of personality among Chinese children aged from 3 to 12]. Unpublished report, Liaoning Normal University, Dalian, China.

Zhuo, M. H. (2008). 2–9 sui er tong qing xu li jie neng li de fa zhan yan jiu [Study of emotional understanding ability among Chinese children aged from 2 to 9]. Unpublished report, Zhejiang University, Hangzhou, China.