Springer Proceedings in Mathematics & Statistics

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

Quantitative Psychology

The 87th Annual Meeting of the Psychometric Society, Bologna, Italy, 2022

Springer

Springer Proceedings in Mathematics & Statistics

Volume 422

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

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

Quantitative Psychology

The 87th Annual Meeting of the Psychometric Society, Bologna, Italy, 2022

Springer

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

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

Jorge González
Facultad de Matemáticas, and Millennium Nucleus on Intergenerational Mobility: From Modelling to Policy (MOVI)
Pontificia Universidad Católica
Santiago, Chile

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

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

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-031-27780-1 ISBN 978-3-031-27781-8 (eBook)
https://doi.org/10.1007/978-3-031-27781-8

Mathematics Subject Classification: 62-06, 62P15

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

Preface

Most countries decided to open up their society again after the pandemic in the beginning of 2022. This meant that we could finally meet in person again, and the 87th annual meeting of the Psychometric Society was held in Bologna, during July 11–15, 2022. Some of the presentations given at that meeting are included in this volume. There were 325 abstracts submitted (including 198 oral presentations, 57 posters, and 144 papers in organized symposia). The meeting attracted 388 participants, 80 of whom also participated in short course pre-conference workshops. There were three keynote presentations, six invited presentations, three spotlight speaker presentations, two dissertation award presentations, one early career award presentations, and one career award presentation.

This will be the eleventh time that Springer publishes the proceedings volume, from the annual meeting of the Psychometric Society. This volume is important as it allows presenters at the annual meeting to spread their ideas quickly to the wider research community, while still undergoing a thorough review process. The previous ten volumes of the IMPS proceedings were received successfully, and we expect these proceedings to be successful as well.

We asked the authors to use their presentations at the Bologna meeting as the basis of their chapters. The authors also had the possibility to extend their chapters with new ideas or additional information. The result is a selection of 32 state-of-the-art chapters addressing several different aspects of psychometrics. The content of the chapters includes, but are not limited to, item response models, structural equation models, missing values, test equating, cognitive diagnostic models, and different kind of applications.

Umeå, Sweden	Marie Wiberg
Amsterdam, The Netherlands	Dylan Molenaar
Santiago, Chile	Jorge González
Madison, WI, USA	Jee-Seon Kim
Montreal, QC, Canada	Heungsun Hwang

Contents

Factors Affecting Efficiency of Interrater Reliability Estimates from Planned Missing Data Designs on a Fixed Budget

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

Abstract Estimating interrater reliability (IRR) requires each of multiple subjects to be observed by multiple raters. Recruiting subjects and raters may be problematic: There may be few available, it may be costly to compensate subjects or to train raters, and participating in an observational study may be burdensome. Planned missing observational designs, in which raters vary across subjects, may accommodate these problems, but little guidance is available about how to optimize a planned missing observational design when estimating IRR. In this study, we used Monte Carlo simulations to optimize an observational design to estimate intraclass correlation coefficients (ICCs), which are very flexible IRR estimators that allow missing observations. We concluded that, given a fixed total number of ratings, the point and credibility estimates of ICCs can be optimized by means of (approximately) continuous measurement scales and assigning small teams of raters to subgroups of subjects. Also, less substantial differences between raters resulted in more efficient IRR estimates. These results highlight the importance of well-designed observational designs and proper training on an observational protocol to avoid substantial differences between raters.

Keywords Interrater reliability · Intraclass correlation · Generalizability theory · Planned missing data · Observational design

L. A. van der Ark (✉) · T. D. Jorgensen
Universiteit van Amsterdam, Amsterdam, the Netherlands
e-mail: L.A.vanderArk@uva.nl; T.D.Jorgensen@uva.nl

D. ten Hove
Universiteit van Amsterdam, Amsterdam, the Netherlands
Vrije Universiteit Amsterdam, Amsterdam, the Netherlands
e-mail: D.ten.Hove@vu.nl

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_1

1 Introduction

This chapter provides evidence we gathered to plan a complex observational study for the Netherlands Ministry of Justice and Security to estimate the interrater reliability (IRR) of the National Instrument of the Juvenile Criminal Justice System, which is known by its acronym LIJ (pronounced like the English word "lie"; Van der Put et al., 2011). The LIJ is used to predict the risk of recidivism and to identify protective factors and risk factors of all minors who are suspect of a criminal case. Completing the LIJ includes an officer of the Netherlands Child Care and Protection Board (rater) separately interviewing both the juvenile (subject) and at least one caretaker to obtain answers to almost 200 questions. This procedure typically takes several workdays spent on reading the police files, conducting the two interviews, and obtaining additional information from and verifying information with, for example, social workers or teachers (see Van der Ark et al., 2018, with a summary in English on p. 5).

Estimating IRR requires that each subject is assessed by multiple raters. Three main challenges complicated investigating the IRR of the LIJ. First, a lack of time. The officers—who also have other important job responsibilities—lacked the time to obtain multiple ratings of the same juveniles. Second, the pool of raters and subjects to choose from was limited. Ecologically valid ratings require raters who are real officers and subjects who are real juveniles within the justice system. Third, recording interviews with the juveniles would be too ethically risky, but raters were required to make observations at the same time and location. The LIJ was administered on 18 different locations in The Netherlands. Obtaining multiple ratings of the same subjects was thus complicated by constraints on time and resources. From a pragmatic perspective, each juvenile was preferably assessed by a minimal number of raters from a local team.

Sampling few raters minimizes the burden on subjects and raters, but maximizes sampling variability of IRR estimates. Because the stakes were high for the juvenile delinquents, precise IRR estimates were required. Planned missing observational designs in which the raters vary across subjects enable using a larger sample of raters while keeping the burden on individual raters stable. Guidance in optimizing such a planned missing observational design to yield precise IRR estimates is currently lacking. In this chapter, we therefore discuss a simulation study that aimed to yield IRR estimates with maximal precision while minimizing the burden on raters.

1.1 Intraclass Correlation Coefficients

IRR coefficients that can accommodate incomplete data are rare (e.g., Krippendorff's α; Hayes & Krippendorff, 2007), but an advantageous choice is the intraclass correlation coefficient (ICC), which has long been used to quantify IRR (Bartko, 1966; Fleiss & Cohen, 1973; Shrout & Fleiss, 1979). A family of ICC coefficients can be derived from the broad framework of generalizability theory

(GT; Cronbach et al., 1963; Brennan, 2001), which was developed for normally distributed variables (McGraw & Wong, 1996) and can be calculated from variance components estimable using a linear mixed model (Jiang, 2018):

$$Y_{sr} = \mu + \mu_s + \mu_r + \mu_{sr}, \tag{1}$$

where Y_{sr} is rating of subject s by rater r, μ is the overall mean rating, μ_s and μ_r are main subject and rater effects, respectively, and μ_{sr} is the subject × rater interaction (confounded with any other source of measurement error). Assuming independent effects with means of zero, the orthogonal variance components sum to the total variance of Y_{sr}:

$$\sigma_Y^2 = \sigma_s^2 + \sigma_r^2 + \sigma_{sr}^2. \tag{2}$$

An ICC quantifying absolute agreement among raters expresses the variance between subjects relative to all sources of variance (McGraw & Wong, 1996):

$$\text{ICC(A, 1)} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_r^2 + \sigma_{sr}^2}, \tag{3}$$

and can be interpreted as the degree to which subjects' absolute scores can be generalized over raters (relevant when evaluating whether a subject meets an absolute criterion; Vispoel et al., 2018; Ten Hove et al., 2023). If, in practice, judgments would be made by averaging scores across $k > 1$ raters, reliability would be increased by reducing rater-related error proportional to k; that is, by dividing the rater-related variance components in the denominator of Equation (3) by the number of raters $\frac{\sigma_r^2 + \sigma_{sr}^2}{k}$.

Equation (1) can be extended by relying on the latent response variable (LRV) interpretation of a probit model (Agresti, 2007). Assuming an observed outcome X—measured using a discrete scale with categories[1] $c = 0, \ldots, C$—is a crude indicator of an underlying continuum Y:

$$X_{sr} = c \text{ if } \tau_c < Y_{sr} \leq \tau_{c+1}, \tag{4}$$

a standard linear-model interpretation is applicable to the LRV Y_{sr}, under an identification constraint that the residual variance $\sigma_{sr}^2 = 1$. An advantage is that ICCs can be compared across studies that used different response scales, such as binary vs. 5- or 7-point Likert scales (Zumbo et al., 2007; Vispoel et al., 2019). The LRV approach has recently been proposed for generalizability coefficients (of which ICCs are a special case; Ten Hove et al., 2023; Vispoel et al., 2018) using structural equation modeling (SEM; Vispoel et al., 2019; Ark, 2015), which Jorgensen (2021)

[1] There are actually $C + 2$ thresholds, but the lowest and highest thresholds are fixed by definition to be the lowest and highest possible scores in the Y distribution; because the normal distribution is unbounded, $\tau_0 = -\infty$ and $\tau_{C+1} = +\infty$.

showed could be problematic for sparse data from planned missing data (PMD) designs. The current study investigates a generalized linear mixed model (GLMM) with a (cumulative) probit link function for ordinal outcomes, which can estimate variance components from a crossed design even with incomplete data.

1.2 Planned Missing Data

PMD designs were conceived to reduce participant burden in large scale surveys (Graham et al., 1996). We introduce some terminology to facilitate discussing PMD designs in the context of multirater studies. Regardless of whether the limits are monetary, we refer to the total number of ratings (N_{Ratings}) as the *budget*. A fixed budget could be limited not only by time and monetary constraints but also by the numbers of available subjects and raters. We further define *workload* as the number of subjects per rater ($N_{S/R}$) and *team size* as the number of raters per subject ($N_{R/S}$). How the budget is allocated depends on a number of features, listed in Table 1. The overall number of subjects and raters are represented by N_S and N_R. Different (sub)samples from the pool of raters might be assigned to each subject, and different (sub)samples of the subject pool may be assigned to each rater. In a fully crossed two-way design with complete data, $N_{\text{Ratings}} = N_S \times N_R$ because the number of subjects assigned to each rater ($N_{S/R} = N_S$) is the entire subject pool; likewise, the number of raters assigned to each subject ($N_{R/S} = N_R$) is the entire rater pool. Incomplete designs are still crossed but do not assign each rater to every subject (Ten Hove et al., 2023). Putka et al. (2008) referred to *ill-structured measurement designs* when assignment was not systematic or optimal, but thoughtfully deployed PMD designs can be economically advantageous in multirater studies with expensive or time-consuming observational protocols (e.g., Vial et al., 2019; Zee et al., 2020; Yuen et al., 2020).

Randomly or systematically assigning a $N_{S/R}$ subset of the subject pool to be observed by each rater (or vice versa: a $N_{R/S}$ subset of the rater pool is assigned to observe each subject) has been shown to improve accuracy of estimated variance components used to calculate an ICC to represent IRR (Ten Hove et al., 2020, 2021). For example, Yuen et al. (2020) randomly assigned a team of two raters to each subject in a staggered fashion that maximized the overlap among raters (i.e., each possible pair of raters rated the same subject at least once). If only $N_R = 2$ raters had observed all $N_S = 29$ subjects, each rater would have a workload of $N_{S/R} = 29$. Instead, each of $N_R = 6$ raters had a substantially lower workload of only $N_{S/R} = 9$ or 10. Thus, given a fixed budget ($N_{\text{Ratings}} = 58$), sampling the same team size ($N_{R/S} = 2$) from a larger pool of $N_R = 6$ raters reduced the workload by $\frac{N_{R/S}}{N_R} = 1/3$.

The simulation study described next was designed to decide how IRR of the LIJ could be most efficiently estimated under budget constraints. The results led Van der Ark et al. (2018) to evaluate the LIJ by assigning teams of $N_R = 4$ raters to evaluate $N_{S/R} = 2$ subjects each.

Table 1 Trade-off among rater-pool size, subject-pool size, team size, and workload given a fixed budget

Reduction	Consequence
A *smaller* pool of subjects[a]	... requires assigning *more* raters per subject[c] (larger teams)
A *smaller* pool of raters[b]	... requires assigning *more* subjects per rater[d] (greater workload)
Assigning *fewer* raters per subject[c] (smaller teams)	... requires a *larger* pool of subjects[a]
Assigning *fewer* subjects per rater[d] (lighter workload)	... requires a *larger* pool of raters[b]

Note: Budget = total number of ratings ($N_{\text{Ratings}} = N_R \times N_{S/R} = N_S \times N_{R/S}$), assuming equal team sizes across subjects and equal workload across raters. When using a block design (i.e., nonoverlapping teams), additionally useful design features can be derived, albeit redundant with the features above: block size = $N_{S/R} \times N_{R/S}$ and $N_{\text{Blocks}} = \frac{N_{\text{Ratings}}}{\text{block size}}$

[a] Subject pool: $N_S = N_R \times \frac{N_{S/R}}{N_{R/S}}$

[b] Rater pool: $N_R = N_S \times \frac{N_{R/S}}{N_{S/R}}$

[c] Team size: $N_{R/S} = N_{S/R} \times \frac{N_R}{N_S}$

[d] Workload: $N_{S/R} = N_{R/S} \times \frac{N_S}{N_R}$

2 Method

To develop an observational design for estimating the IRR of the scales and items of the LIJ, we conducted a set of Monte Carlo simulations. We provide our R syntax for replicating our simulation on the Open Science Framework (OSF[2]).

2.1 Data Generation

The two-way model in Eq. (1) was used to generate normal random effects for all conditions, with $\mu = 0$ for the grand mean and all random-effect means, $\sigma_s^2 = 0.70$, $\sigma_r^2 = 0.15$, and $\sigma_{sr}^2 = 0.25$. These population variances implied a population ICC(A,1) = 0.636, denoted ρ. For ordinal conditions, thresholds $\tau_1 = -0.5$ and $\tau_2 = 0.5$ were used to discretize the continuous data into $C = 3$ categories. To keep the generated data comparable across conditions, the data were always generated from a fully crossed design for a given N_S and N_R. Then, missing data patterns were imposed to yield a certain number of complete-data "blocks" (i.e., $N_{\text{Blocks}} = N_{R/S} \times N_{S/R}$) that yielded a fixed budget of $N_{\text{Ratings}} = 384$.

[2] Supplemental online materials available at https://osf.io/g5hvs/.

2.1.1 Core Design Factors

The design factors we had most control over were workload and team size given a fixed budget, and we planned to estimate ICCs for continuous, ordinal, and binary items. So our core factors were team size ($N_{R/S} = 2, 4,$ or 8), workload ($N_{S/R} = 1, 2, 4,$ or 8), and model used for generating and analyzing data (linear or probit for continuous or discrete data, respectively), yielding $3 \times 4 \times 2 = 24$ conditions. The proportion of missing data in two-way designs (i.e., $N_{S/R} > 1$) varied from 83.33% (in conditions with the largest blocks) to 98.96% (with the smallest workload $N_{S/R} = 2$ and team size $N_{R/S} = 2$, requiring the largest N_R and N_S). When $N_{S/R} = 1$, there is no "missing-data problem" because raters are nested in (rather than crossed with) subjects. For $N_{S/R} = 1$, we used an one-way model by removing μ_r from Eq. (1) and its variance component σ_r^2 from Eq. (2) because when raters are nested in subjects, μ_r is confounded with the rater $\times$ subject interaction. Thus, Eq. (3) still represents ICC(A,1).

2.1.2 Additional Design Factors

In the results section, we also describe two follow-up studies in which we varied two additional factors: The magnitude of reliability, and random versus block assignment of raters to subjects. We fully crossed these design factors with the core conditions described above, but did not cross these with each other. The results are useful for planning missing observational designs for IRR. Additional manipulations are available in the R scripts provided with the online supplementary materials.

2.2 Analysis

We used Markov chain Monte Carlo (MCMC) estimation with uninformative priors, implemented in the Stan software (Carpenter et al., 2017), for each of 2000 replications within each condition. We saved the posterior mean (denoted $\hat{\rho}$) as an estimate of ρ, as well as the central 95% Bayesian credible interval (BCI) limits. In each condition, we evaluated accuracy of posterior means as point estimates by calculating the relative parameter bias, which is the difference between a condition's average estimate (denoted $\bar{\rho}$) and ρ, divided by ρ: $\frac{\bar{\rho}-\rho}{\rho}$. We evaluated accuracy of BCIs by calculating 95% coverage rates (i.e., proportion of replications whose intervals captured ρ) in each condition. Finally, we evaluated precision (our primary criterion for choosing an optimal design for the LIJ evaluation) of the estimates by calculating the average width of 95% BCIs in each condition.

We investigate the effects of design factors on bias and precision using fully factorial linear regression models (ANOVA) and on the coverage using fully factorial binary logistic regression models (analysis of deviance).

3 Results

For brevity, we report only medium and larger effects (i.e., Monte Carlo design factor accounts for $\eta_p^2 > 6\%$ of variance, holding other effects constant) on bias or precision, or 6% of deviance in coverage (analogous to McFadden's[3] pseudo-R^2). More extensive results are provided on the OSF.

3.1 Core Conditions

Table 2 shows results across core conditions, and Figs. 1 and 2 show the width of the 95% BCIs across conditions with continuous and ordinal responses, respectively.

Table 2 Relative bias, CI coverage, and CI width across core conditions

Scale	Workload ($N_{S/R}$)	Team size ($N_{R/S}$)	Bias	Coverage	Width
Continuous	1	2	−0.017	0.946	0.173
	1	4	−0.011	0.944	0.174
	1	8	−0.010	0.951	0.213
	2	2	−0.019	0.945	0.174
	2	4	−0.013	0.950	0.176
	2	8	−0.014	0.940	0.216
	4	2	−0.022	0.934	0.176
	4	4	−0.016	0.951	0.181
	4	8	−0.014	0.951	0.218
	8	2	−0.024	0.950	0.183
	8	4	−0.021	0.944	0.189
	8	8	−0.019	0.946	0.225
Ordinal	1	2	0.011	0.941	0.226
	1	4	0.012	0.938	0.211
	1	8	0.011	0.941	0.241
	2	2	0.005	0.934	0.226
	2	4	0.009	0.946	0.214
	2	8	0.006	0.935	0.244
	4	2	0.007	0.941	0.226
	4	4	0.006	0.946	0.217
	4	8	0.007	0.938	0.246
	8	2	0.004	0.956	0.229
	8	4	0.003	0.942	0.224
	8	8	0.002	0.942	0.252

Note: Accuracy represented by bias (expressed as a proportion of the true $\rho = 0.636$) and 95% CI coverage. CI width represents precision

[3] Find descriptions of several types of pseudo-R^2 for logistic regression here: https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/.

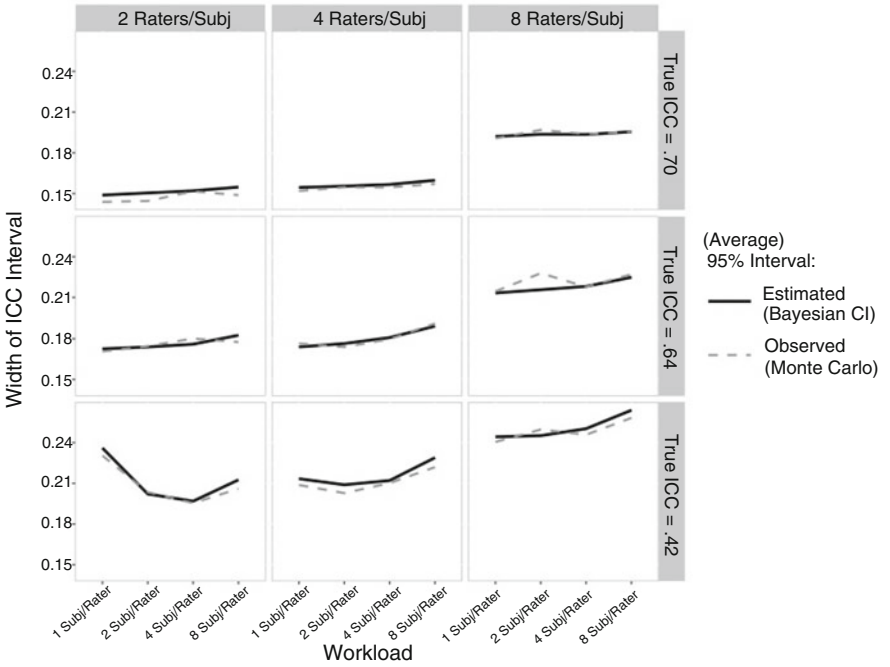


Fig. 1 Width of 95% BCI for ICC(A,1) under different conditions for continuous data

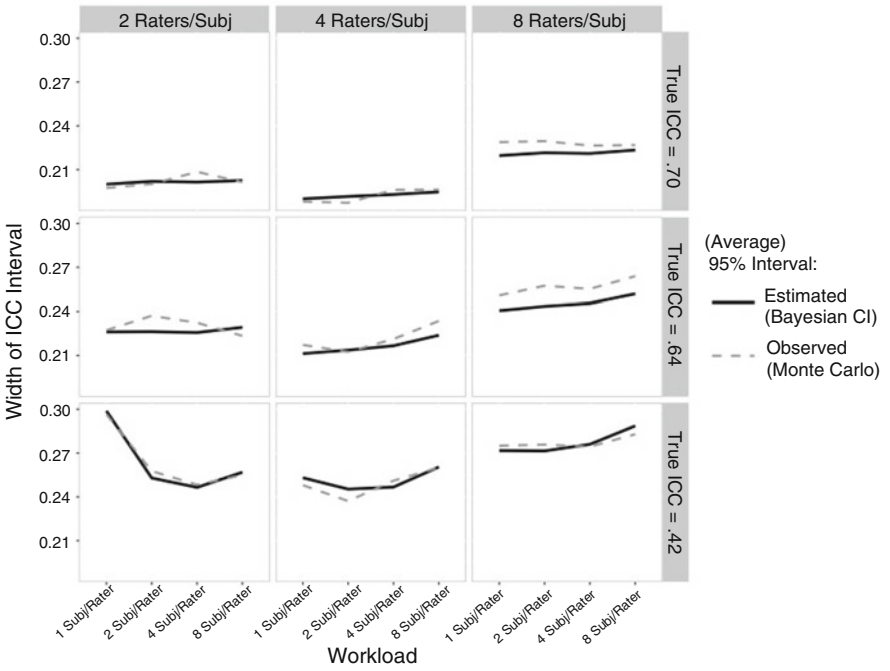


Fig. 2 Width of 95% BCI for ICC(A,1) under different conditions for ordinal data

The width of empirical intervals (i.e., distance between 2.5% and 97.5% quantiles in the distribution of posterior means across Monte Carlo replications) are included for comparison; similarity with BCIs indicates accurate estimates of uncertainty.

Bias was negligible across conditions ($M_{\text{bias}} = -0.01$, $SD = 0.01$), and no design factors explained more than 0.05% of variance in bias. Coverage was nominal across conditions ($M_{\text{cov}} = 0.94$, $SD = 0.01$), and no design factors explained more than 0.05% of deviance in coverage. Precision was substantially affected only by the scale (continuous or ordinal: $\eta^2_{\text{p}} = 14.74\%$) and team size ($\eta^2_{\text{p}} = 11.38\%$). ICC(A,1) was more precisely estimated for continuous data (95% BCI width: $M_{\text{width}} = 0.19$, $SD = 0.03$) than for ordinal data ($M_{\text{width}} = 0.23$, $SD = 0.02$). ICC(A,1) was more precisely estimated using teams of $N_{R/S} = 2$ ($M_{\text{width}} = 0.20$, $SD = 0.03$) or $N_{R/S} = 4$ ($M_{\text{width}} = 0.20$, $SD = 0.03$) than for teams of $N_{R/S} = 8$ ($M_{\text{width}} = 0.23$, $SD = 0.02$).

An explanation of why smaller teams yielded more precise estimates may be that—holding other factors constant—assigning smaller teams ($N_{R/S}$) maximizes N_S (Table 1). Because σ^2_s should be expected to be the largest component of an ICC in practice (e.g., for even a modest IRR ≥ 0.50), a more efficiently estimated σ^2_s could lead to a more efficiently estimated ρ. The next simulation additionally varied the amount of rater error, illuminating this explanation.

3.2 Magnitude of ICC

Rater variance was fixed to $\sigma^2_r = 0.25$ in the core conditions, implying a modest $\rho = 0.636$. Because we expected ICCs to vary across LIJ items, we added conditions with more rater variance ($\sigma^2_r = 0.70$, implying lower IRR: $\rho = 0.42$) and with less rater variance ($\sigma^2_r = 0.05$, implying higher IRR: $\rho = 0.70$). Extending the 24 core conditions by varying $\sigma^2_r = 0.05$, 0.25, or 0.70 yielded 72 conditions.

Bias was still negligible across conditions ($M_{\text{bias}} = -0.002$, $SD = 0.014$) and no design factors explained more than 0.03% of variance in bias. Coverage also still was nominal across conditions ($M_{\text{cov}} = 0.94$, $SD = 0.01$), and no design factors explained more than 0.01% of deviance in coverage. Efficiency was substantially affected by ρ, which explained $\eta^2_{\text{p}} = 15.64\%$ of the variability in BCI width, whereas the influential factors from the core conditions explained only $\eta^2_{\text{p}} = 5\%$ of the variability in BCI width, holding other factors constant in this extended design. Figures 1 and 2 show results for continuous and ordinal data, respectively. The higher IRR was in the population, the more precisely it was estimated ($\rho = 0.70$: $M_{\text{width}} = 0.19$, $SD = 0.03$; $\rho = 0.64$: $M_{\text{width}} = 0.21$, $SD = 0.03$; $\rho = 0.42$: $M_{\text{width}} = 0.25$, $SD = 0.03$). This was consistent with our explanation for why smaller teams yielded more precise estimates under a fixed budget; it is not a general rule that fewer raters (per subject) yield more precision (Ten Hove et al., 2021).

3.3 Overlapping Teams

In the core conditions, we imposed a missing-data structure that mimicked the blocks assigned in the LIJ study. We compared this to unstructured random assignment by randomly deleting all but $N_{R/S}$ ratings for each subject. This strategy meant that the workload could vary across raters, with an average (rather than fixed) workload of $N_{S/R}$. This design is comparable to Yuen et al. (2020), who designed a balanced workload across raters (i.e., fixed $N_{S/R}$). Because overlapping teams implies a two-way design, we omitted the $N_{S/R} = 1$ conditions. Thus, this study had a 3 ($N_{R/S} = 2$, 4, or 8) × 3 ($N_{S/R} = 2$, 4, or 8) × 2 (scale) × 2 (teams overlap or not) design with 36 conditions.

Bias was still negligible across conditions ($M_{\text{bias}} = -0.005$, $SD = 0.013$) and no design factors explained more than 0.11% of variance in bias. Coverage also still was nominal across conditions ($M_{\text{cov}} = 0.94$, $SD = 0.01$), and no design factors explained more than 0.04% of deviance in coverage. Precision was also not substantially affected by overlapping teams; the main and all higher-order effects combined only explained $\eta^2_{\text{p}} = 2.74\%$ of the additional variability in BCI width beyond the core design.

4 Discussion

This chapter provided evidence from Monte Carlo simulations demonstrating how beneficial planned missing observational designs can be for expensive, time-consuming multirater studies. Results showed that MCMC estimation of MLMs and GLMMs can provide accurate point and interval estimates of ICCs across a variety of population values, scales of measurement, and planned missing observational designs, even when the vast majority of observations of a conventional (fully crossed) two-way design are missing. Bias and coverage appeared stable across the selected design factors but we showed that the precision of ICC estimates can be maximized by using more (approximately) continuous scales of measurement and allocating smaller teams of raters to subjects. These results highlight the importance of well designed observational designs. In addition, less rater error also improved efficiency, which highlights the importance of proper training on an observational protocol to avoid substantial differences between raters.

4.1 Advice for Sample-Size Planning

In practice, researchers must weigh the costs of different design features to choose the best design for their situation (e.g., Are raters or subjects more expensive? Does the gain in efficiency warrant the additional effort?), and accounting for such costs

was not explored in our simulation studies. Certain design features might also be more difficult to control than others. Holding other features constant, smaller teams and lighter workloads might require larger pools of subjects and raters, respectively, either of which might be infeasible.

In the LIJ study, for example, the number of available juveniles turned out to be quite limited. Furthermore, the greatest cost was the burden on each rater, so workload was of primary concern in the design. With a fixed budget, minimizing team size to improve precision would have been coincident with maximizing N_S, which was not feasible. However, smaller teams (larger N_S) only improved precision by a few decimal places, and workload had no discernible effect on precision, so we felt justified advising the ministry to assign fewer subjects to larger teams for LIJ data collection.

Overlapping raters does not seem to have any (dis)advantage, so researchers can feel free to randomly assign raters to subjects using whichever algorithm best fits their needs. Overlapping raters might be more feasible if the ratings need not be conducted at a fixed time point; for example, if the subjects have been recorded, or if the observation is made on objects (like critics judging artwork or experts evaluating the face validity of a measurement instrument). Systematic overlap is not necessary, but might be more desirable to ensure balanced workload across raters (see Yuen et al., 2020). In contrast, random assignment to blocks (within which subjects and raters are fully crossed) would be more feasible when live observations of the same event must be made at the same time, as in the LIJ evaluation.

We conducted these simulations for a specific setting (evaluating IRR of the LIJ), and showed that some general design factors can improve the efficiency of ICC estimates. We hope that our example helps other researchers make such decisions, but future research is needed to provide advice for other scenarios that have different priorities for working within a budget.

Appendix

Four annotated R functions to generate sample data from a random-effects model. The first function generates (approximately) normally distributed data, the second discretizes those data with thresholds to make ordinal data. The third function imposes missing-data patterns consistent with the design factors. The fourth function transforms the data from wide to long format with respect to raters.

```
## function to simulate a full matrix of all possible continuous ratings.
## To limit time for Kripp's bootstrap, scale and round to closest 1/3.
simcon <- function(subj  = 0.7,  # variance of subject effect
                   rater = 0.15, # variance of rater effect
                   error = 0.25, # variance of measurement error
                   nS = 16, # numbers of subjects and raters
                   nR = 32) {
  subj.effect  <- rnorm(nS, 0, sqrt(subj))
  rater.effect <- rnorm(nR, 0,sqrt(rater))
  error        <- rnorm(nS*nR, 0, sqrt(error))
  ratings <- matrix(error, nrow = nS, ncol = nR)
```

```
  for (RR in 1:nrow(ratings)) {
    ratings[RR, ] <- ratings[RR, ] + subj.effect[RR]
  }
  for (CC in 1:ncol(ratings)) {
    ratings[ , CC] <- ratings[ , CC] + rater.effect[CC]
  }
  matrix(round(scale(as.numeric(ratings))[,1] * 3) / 3, # 19 categories between
      -3:3
         nrow = nS, ncol = nR)
}
## Test function
# simcon()

## function to generate ordinal data by applying 2 thresholds to continuous
     data
simord <- function(x,  # matrix output by simcon
                   threshold1 = -0.5, threshold2 = 0.5) {
  ratings1 <- x > threshold1
  ratings2 <- x > threshold2
  ratings1 + ratings2
}
## test it
# simord(simcon())


## function to impose missing data patterns on data from simcon() or simord().
pokeHoles <- function(data, random = TRUE, RpS = 4, SpR = 2) {
  nS <- nrow(data)
  nR <- ncol(data)
  if (SpR == 1L) random <- FALSE # SpR == 1L implies independence, no overlap
  ## If random, sample RpS columns (raters) within each row (subject).
  if (random) {
    for (RR in 1:nS) {
      keep <- sample(1:nR, size = RpS)
      data[RR, -keep] <- NA
    }
  } else {
    ## If fixed, create independent blocks of raters with same subjects.
    nProjectjes <- nS / SpR # how many blocks of complete observations
    if (nProjectjes != nR/RpS) stop('Design numbers inconsistent with overall',
                                    ' count of raters or subjects.')
    obsMat <- matrix(TRUE, nrow = SpR, ncol = RpS)
    missMat <- !kronecker(diag(nProjectjes),  obsMat)
    data[missMat] <- NA
  }
  data
}
## test it
# pokeHoles(simcon())
# pokeHoles(simcon(), random = FALSE)


## function to transform matrix of ratings (pokeHoles output) from wide to long
## format.  If ratings are ordinal, convert to ordered factor.
trans <- function(x, ordered = FALSE) {
  library(reshape2)
  long <- melt(x)
  names(long) <- c("subject", "rater", "rating")
  long <- long[!is.na(long$rating), ]
  if (ordered) long$rating <- ordered(long$rating)
  long
}
## test it
# trans(pokeHoles(simord(simcon())))
```

Annotated Stan (Carpenter et al., 2017) syntax to fit a cross-classified hierarchical linear model to the normally distributed two-way data in order to quantify IRR by estimating ICC(A,1) with MCMC-estimated variance components.

```
data {
  int N;           // number of ratings
  int nS;          // number of subjects
  int nR;          // number of raters
  int sID[N];
  int rID[N];
  real Rating[N];
  real rangeRatings;
}
parameters {
  real Intercept;
  vector[nS] eS;  // Subject effects (deviation of true-score from mean rating)
  vector[nR] eR;  // Rater effects
  real<lower=0,upper=rangeRatings/2> sigmaS;  // SD of subject effects
  real<lower=0,upper=rangeRatings/2> sigmaR;  // SD of rater effects
  real<lower=0,upper=rangeRatings/2> sigmaE;  // interaction + residual SD
}
model {
  real mu[N];
  // Priors
  Intercept ~ normal(0, rangeRatings/2);
  eS ~ normal(0, sigmaS);
  eR ~ normal(0, sigmaR);
  //sigmaS ~
  // Fixed effects
  for (n in 1:N)  mu[n] = Intercept + eS[sID[n]] + eR[rID[n]];
  // Likelihood
  Rating ~ normal(mu, sigmaE);
}
generated quantities {
  real icc;
  icc = sigmaS*sigmaS / (sigmaS*sigmaS + sigmaR*sigmaR + sigmaE*sigmaE);
}
```

Additional *.stan files for other models (1-way data, ordinal data) can be found on the OSF, along with the R scripts to generate data, fit the model using the R package `rstan`, run the simulation, and compile and analyze simulation results: https://osf.io/g5hvs/

References

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Wiley.

Ark, T. K. (2015). *Ordinal generalizability theory using an underlying latent variable framework*. Ph.D Thesis, University of British Columbia, Vancouver, BC. https://doi.org/10.14288/1.0166304

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports, 19*(1), 3–11. https://doi.org/10.2466/pr0.1966.19.1.3

Brennan, R. L. (2001). *Generalizability theory*. Springer. https://doi.org/10.1007/978-1-4757-3456-0

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*(2), 137–163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

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*(3), 613–619. https://doi.org/10.1177/001316447303300309

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research, 31*(2), 197–218. https://doi.org/10.1207/s15327906mbr3102_3

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77–89. https://doi.org/10.1080/19312450709336664

Jiang, Z. (2018). Using the linear mixed-effect model framework to estimate generalizability variance components in R: A `lme4` package application. *Methodology, 14*(3), 133–142. https://doi.org/10.3758/s13428-017-0986-3

Jorgensen, T. D. (2021). How to estimate absolute-error components in structural equation models of generalizability theory. *Psych, 3*(2), 113–133. https://doi.org/10.3390/psych3020011

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology, 93*(5), 959–981. https://doi.org/10.1037/0021-9010.93.5.959

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2020). Comparing hyperprior distributions to estimate variance components for interrater reliability coefficients. In M. Wiberg, D. Molenaar, J. González, U. Böckenholt, & J.-S. Kim (Eds.), *Quantitative psychology: The 84th annual meeting of the Psychometric Society, Santiago, Chile, 2019* (pp. 79–93). Springer. https://doi.org/10.1007/978-3-030-43469-4_7

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2021). Interrater reliability for multilevel data: A generalizability theory approach. *Psychological Methods, 27*(4), 650–666.

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2023). Updated guidelines on selecting an intraclass correlation coefficient for interrater reliability, with applications to incomplete observational designs. *Psychological Methods*. https://doi.org/10.1037/met0000516

Van der Ark, L. A., Van Leeuwen, J. L., & Jorgensen, T. D. (2018). Interbeoordelaarsbetrouwbaarheid LIJ: Onderzoek naar de interbeoordelaarsbetrouwbaarheid van het landelijk instrumentarium jeugdstrafrechtketen [Interrater reliability LIJ: Research on the interrater reliability of the national instrument of the juvenile criminal justice system]. Technical Report, Wetenschappelijk Onderzoek- en Documentatiecentrum, The Hague, the Netherlands. Retrieved from http://hdl.handle.net/20.500.12832/2267.

Van der Put, C., Spanjaard, H., Van Domburgh, L., Doreleijers, T., Lodewijks, H., Ferwerda, H., Bolt, R., & Stams, G. J. (2011). Ontwikkeling van het landelijke instrumentarium jeugdstrafrechtketen (LIJ) [Development of the national instrument of the juvenile criminal justice system]. *Kind & Adolescent Praktijk, 10*(2), 76–83. https://doi.org/10.1007/s12454-011-0021-2

Vial, A., Assink, M., Stams, G. J. J. M., & Van der Put, C. (2019). Safety and risk assessment in child welfare: A reliability study using multiple measures. *Journal of Child and Family Studies, 28*, 3533–3544. https://doi.org/10.1007/s10826-019-01536-z

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods, 23*(1), 1–26. https://doi.org/10.1037/met0000107

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2019). Using generalizability theory with continuous latent response variables. *Psychological Methods, 24*(2), 153–178. https://doi.org/10.1037/met0000177

Yuen, J. K., Kelley, A. S., Gelfman, L. P., Lindenberger, E. E., Smith, C. B., Arnold, R. M., Calton, B., Schell, J., & Berns, S. H. (2020). Development and validation of the ACP-CAT for assessing the quality of advance care planning communication. *Journal of Pain and Symptom Management, 59*(1), 1–8. https://doi.org/10.1016/j.jpainsymman.2019.09.001

Zee, M., Rudasill, K. M., & Roorda, D. L. (2020). "Draw me a picture": Student–teacher relationship drawings by children displaying externalizing, internalizing, or prosocial behavior. *The Elementary School Journal, 120*(4), 636–666. https://doi.org/10.1086/708661

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods, 6*(1), 21–29. https://doi.org/10.22237/jmasm/1177992180

Concordance for Large-Scale Assessments

Liqun Yin, Matthias Von Davier, Lale Khorramdel, Ji Yoon Jung, and Pierre Foy

Abstract Interest has grown recently in linking national or regional assessments to international large-scale assessments. However, commonly used equating and linking methods are not defensible for such purposes as they would make unrealistic assumptions such as construct equivalency and error-free measurement, and usually only provide a point to point projection. This paper introduces a new approach for score projection by constructing an enhanced concordance table between two large-scale assessments with one source test and one target test. Specifically, the proposed method employs predictive mean matching method to find a set of donors with the smallest distances to the predicted mean generated by an imputation model on the source test for each concordance level within the identified score range. Both the means and standard deviations of donors' plausible values on the target test are utilized to construct a concordance table between the two tests. This approach not only ensures the score uncertainty due to measurement error and imperfect correlation between tests are appropriately taken into account, but also avoids complex statistical functional forms and linearity assumption. The robustness of the new approach is demonstrated by a linking study to relate a regional assessment to TIMSS and PIRLS international long-standing large-scale assessments, where students take both the source and the target tests. Recommendations for educators and researchers to make inferences and interpret the concordance table are also provided.

Keywords Large-scale assessments · Linking and equating · Concordance · TIMSS

L. Yin (✉) · M. Von Davier · L. Khorramdel · J. Y. Jung · P. Foy
Boston College, TIMSS & PIRLS International Study Center, Boston, MA, USA
e-mail: yinld@bc.edu; vondavim@bc.edu; lale.khorramdel@bc.edu; jiyoon.jung@bc.edu; foypi@bc.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_2

1 Introduction

Over the past 20 years, large-scale assessment has become an increasingly more popular field of study in education. Interest has also grown recently in situations where results from national and regional large-scale assessments are to be linked to international large-scale assessments (Hernández-Torrano & Courtney, 2021). A practical situation is using the results of one assessment to estimate the likely score range on another large-scale assessment as if it had been administered. For example, if a country participated in a regional large-scale assessment but is not an international assessment participant yet, researchers and educators may be curious to know what the percentage of the students reaching international benchmarks or proficiency levels would be. Linking a national test to an international test provides the opportunity to locate the outcomes of the national study on an established international scale. Being able to project scores or proficiency levels from different assessments onto one well-known international scale is also very useful and highly desired for benchmarking in educational monitoring.

Various methodologies can be used for linking or equating assessments. There are three broad categories of linkages between tests (Holland & Dorans, 2006; Linn et al., 2009; Mislevy, 1992): equating two tests X and Y, aligning the scales of tests X and Y, predicting/projecting the likely results of test Y from test X. Equating makes strong assumptions about the two test scores to be equated (Lord, 1980) and is hence considered the strongest form of scale linkage. Equating is a statistical procedure that is used to adjust the scores of one test, among a pair of two (essentially parallel) tests targeting the same construct using the same test blueprints. The goal is to create two sets of scores that can be used interchangeably (Kolen & Brennan, 2014). A similar approach to equating, known as scale alignment, can be used to achieve comparability between two tests built to similar specifications but do not meet the parallel forms assumption. Projecting is a concordance approach. A concrete statistical method for linking two measurement scales and calculating a concordance table usually uses equating-like methods (e.g., equipercentile methods) or statistical moderations (e.g., using a series of complex equations to adjust the test scores to have the same mean and standard deviation) to match the scores on two tests that have similar construct but differ in content and/or specifications. Projecting is the least restrictive type of linkage in that it does not assume that the constructs are the same or scores are exchangeable after linking.

2 Linking Large-Scale assessments

Large-scale assessments are tests, usually standardized, or other data collection procedures administered to large numbers of students at the same time. Some of those large-scale assessments require scores for individual students such as state assessments which are used to monitor student performance from year to year.

However, some large-scale assessments are designed with the purpose of reporting results at the group level based on a set of plausible values (PVs) and generally rely on sampling techniques. They often make use of sampling weights and replication methods, resort to item response theory, and utilize item responses and context variables in population models for the calculation of scale scores (e.g., von Davier & Sinharay, 2013). These include international assessments as well as national and regional assessments. Recently, as large-scale assessments with the purpose of reporting results at the group level gain more popularity and publicity than ever before, interest in linking large-scale assessments from national, regional, to international has also grown.

There are a few challenges with linking large-scale assessments in practice. Utilizing a population model that include individual responses from both tests as well as context variables requires data collection designs that are very costly to implement. The construct equivalency assumptions associated with conventional linking methods are typically unrealistic and not defensible in linking regional to international large-scale assessments because the tests being linked measure somewhat different constructs and are constructed in different ways.

Several studies attempted to link test scores from national and international large-scale assessments over a long period of time (Cartwright et al., 2003; Nissen et al., 2015; Jia et al., 2014; Ehmke et al., 2020). Most previous studies used traditional or IRT-based equating-like methods, or alternatively, attempted statistical moderation. In the 2011 NAEP-TIMSS linking study (Jia et al., 2014), the objective of the study was to use states' 2011 NAEP (The National Assessment of Educational Progress in the United States) scores to predict their average TIMSS scores and percentages of students reaching each of the TIMSS international benchmark levels. Two different linking methods were applied, IRT-based calibration linking analysis and statistical moderation. For the calibration linking analysis, the NAEP items were calibrated onto the TIMSS metric by fixing IRT item parameters for the TIMSS items in the specific braided linking booklets to the values from the TIMSS 2011 operational analysis. For statistical moderation, a series of complex equations were utilized to adjust the NEAP scores to have the same mean and standard deviation as TIMSS.

Nissen et al. (2015) also used the IRT-based calibration linking method and an equating method (equipercentile) to link the National Educational Panel Study in Germany to the TIMSS scale. Among the methods used in linking large-scale assessments, the statistical moderation procedures are conceptually and procedurally complex. Equating-like method makes strong assumptions such as construct and reliability equivalency, and that equating functions are invertible. For example, the equipercentile linking uses the same percentile rank across two tests to calculate the expected or predicted score to create a concordance between two tests. The expected score is a concordant or equated score only when the two sets of scores are almost perfectly related. However, regional and international large-scale assessments are usually not very highly correlated, not to speak of measuring directly comparable constructs, since difficulty, construct definitions, and assessments frameworks are not the same.

In this paper, a new approach for score projection is proposed to establish an enhanced concordance table between two large-scale assessments with one source test (X) and one target test (Y). The proposed new method takes the uncertainty of the proficiency estimates on both tests into account and also controls for potential construct differences between the tests. More specifically, it can be conceptually compared to predictive mean matching (PMM; Little, 1988; Rubin, 1986), a customary form of imputation which calculates the predicted value of target variable Y according to the specified imputation model and based on values observed elsewhere, so they are realistic. Note that the model used to generate PVs in large-scale assessments is also a special case of an imputation model (von Davier & Sinharay, 2013). It provides a method for score projection where equating-like methods are not defensible as they would make unrealistic assumptions such as equivalency of constructs and high reliability levels.

3 Technical Procedure for Establishing Concordance Tables

The technical procedures described in this section draw on the statistical principles of population (or conditioning) models used in large-scale assessments (e.g., von Davier et al., 2009). This allows constructing a concordance that incorporates the uncertainty of the projection by utilizing conditional variance estimates. The approach can be described as follows:

The predictive means of source test score θ and target test score ϑ are derived utilizing population models. The expected values given item responses and context data, which provides students' background information is related to achievement such as students' gender and social-economic status, are given by

$$\hat{\vartheta} = E\left(\vartheta | Y_1, \ldots, Y_J, Z_1, \ldots, Z_K\right) \text{ and } \hat{\theta} = E\left(\theta | X_1, \ldots, X_I, Z_1, \ldots, Z_K\right) \tag{1}$$

Focusing on predicting ϑ from θ, the conditional distribution can be constructed for generating imputations even for those cases where only test X is given together with the context variables $Z_1, \ldots, Z_K$. It can be constructed if the imputation models for ϑ and θ can be estimated from a sample, so that the conditional distribution in Eq. (2) can be constructed for generating imputations.

$$P\left(\vartheta | X_1, \ldots, X_I, Z_1, \ldots, Z_K\right) \tag{2}$$

For a concordance, the full population model using individual responses and context variables is often impractical. Practitioners want to use a score on one test to make inferences about the likely score range on another test. This is always projection-based using joint or conditional distributions, and the use of just a point estimate on target test given the source test score would be ignoring the uncertainty of this projected score. Therefore, the approach used here utilizes PVs (obtained from

population models) to account for the uncertainty of the score projection. The observed joint distribution of source and target test latent variable estimates can be used to create a conditional distribution of the target test's latent variable given the source test's variable, $P(\vartheta|\theta)$. Based on a sample of respondents $v = 1, ..., N$, plugging in the posterior means and PVs allows us to approximate this conditional distribution. Instead of constructing the full population model

$$\hat{\vartheta} \sim P\left(\vartheta|X_1, \ldots, X_I, Z_1, \ldots, Z_K\right) \tag{3}$$

an approximate imputation model $P(\vartheta|\theta)$ based on the source and target latent variables only is used and estimated using the two full population models

$$\hat{\vartheta} \sim E\left(\vartheta|Y_1, \ldots, Y_J, Z_1, \ldots, Z_K\right) \text{ and } \hat{\theta} \sim E\left(\theta|X_1, \ldots, X_I, Z_1, \ldots, Z_K\right) \tag{4}$$

to generate an estimate of the conditional distribution

$$P\left(\hat{\vartheta}|\hat{\theta}\right) \approx P\left(\vartheta|\theta\right) \tag{5}$$

Then, the concordance is essentially given by

$$P\left(\hat{\vartheta}|E\left(\theta|X_1, \ldots, X_I, Z_1, \ldots, Z_K\right)\right) \tag{6}$$

and provides a projected distribution on the target test form given a function of the context variable and observed item responses on the source test form.

4 Practical Implementation for Establishing Concordance Tables

To establish a concordance table between two large-scale assessments, 5 main steps are usually involved in practice. The details of these steps are described in this section.

1. Collect data on both tests in a linking sample and estimate conditional distributions

The national test or regional test is usually the source test. The target test is the test to be linked to, usually the international test such as Trends in International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS), or Programme for International Student Assessment (PISA). In a linking study, the test data for the source and the target tests are usually collected at the same time window. After data collection, the next step is to derive the

conditional distributions for individuals through population models of the context and achievement data for the source test and target test, respectively. Then draw sets of PVs (*e.g.*, 5 PVs of each test) separately from the conditional distributions and transform the PVs on the corresponding reporting metrics if it is needed.

2. Identify score range of source test and concordance levels

The concordance score range and levels are identified based on estimated posterior means of the conditional distributions derived from the population models. Using the full range of source test for concordance table is not always practical because outliers may exist. However, the final identified score range for concordance table should cover the majority of data (99% or more). For the concordance levels, or concordance score points, the rule is to include enough levels and to retain as much information as possible but not too trivial. For example, if the standard deviation of reporting metric for the source test is 100, 20 or 10 points apart could be reasonable choices. The choice also depends on the sample size of the linking study available to find donors.

3. Select a set of donors for each concordance level

For each specified concordance level, a PMM method is used to find a set of donors who are the nearest neighbors to each concordance score point. This selection is achieved by selecting a set of (*e.g.*, 5 or 7) smallest absolute differences of students' posterior mean on the source test to each specified concordance level. The posterior mean can be either directly estimated based on the population model identified for generating PVs or, if needed, approximated by simply averaging the generated PVs on the source test.

4. Establish a concordance table between the two tests

All donors' PVs on the target test are utilized to construct the concordance table. The concordance table includes predicted conditional distributions, both predicted means and predicted standard deviations on the target test, given concordance score levels and nearest neighbor donors. Specifically, preliminary concordance tables are created by assigning the mean of all donors' PVs based on the target test data to each corresponding concordance score level as the projected mean within the specified concordance range. The standard deviation of each set of combined donor's PVs based on the target test is the associated uncertainty of the projected mean on the target test.

5. Smoothing and extrapolating

If there are a limited number of samples in the source data, a smoothing procedure such as a simple moving average (*e.g.*, Isnanto, 2011) may be used to better represent the underlying projected conditional means and standard deviations on the target scales in the specified range. To obtain a robust prediction for concordance scores beyond the specified range, where only a very small number of students could be observed, an extrapolation such as Sen's slope estimator or the Thiel-Sen estimator (Sen, 1968) may be used to extrapolate for the concordance score levels

at two ends. The Sen's slope estimator is to find the median of all slopes for all pairs of ordered (ordinal) two variables. Specifically, the median slope for all pairs of ordered concordance score levels and the projected means is used to extrapolate the predicted mean. Similarly, the median slope for all pairs of ordered concordance score levels and the projected standard deviation is used to extrapolate the predicted standard deviation at the two very ends.

5 Example for Establishing Concordance Tables

This section describes the procedures used to construct the International Association for the Evaluation of Educational Achievement (IEA) Rosetta Stone Study concordance table, which provides a projection of the scores on the regional assessment onto the scales of the target assessments, TIMSS and PIRLS.

IEA's Rosetta Stone study is designed to facilitate measuring progress toward the UNESCO Sustainable Development Goal for quality in education and aims at linking different regional assessment programs to TIMSS & PIRLS international long-standing metrics and benchmarks of achievement (IEA website; UNESCO website). The goal is to provide countries who participated in regional assessments but not in TIMSS & PIRLS with information about the proportions of primary school students that have achieved established international proficiency levels in literacy and numeracy for allowing international comparisons. Rosetta Stone Study includes linking different regional large-scale assessments to the TIMSS and PIRLS international long-standing metrics of achievement. In this paper, the Rosetta Stone ERCE (UNESCO's Regional Comparative and Explanatory Study) linking study is used to illustrate the proposed method.

The Rosetta Stone ERCE study has two assessment parts: Rosetta Stone linking booklets, which contain both TIMSS and PIRLS items (items were originally developed for students in grade 4 and were presented in TIMSS 2015 and PIRLS 2016), and the ERCE 6th grade assessment, which tests 6th grade students from Latin American and Caribbean countries in reading and mathematics. The two assessments have similar constructs but are based on different frameworks targeting different populations as defined by their intended focal grade levels. Both assessments include achievement booklets and a set of context questionnaire.

To construct the concordance, the 2019 ERCE assessment was administered to students in the 6th grade together with the Rosetta Stone linking booklets. Students in two countries, 3108 in Colombia and 4716 in Guatemala, participated in the study. For the source tests, ERCE math and reading, items were already calibrated by the ERCE team, which also provided the PVs for ERCE. The IRT scaling and population modeling for the TIMSS & PIRLS linking items were based on those students and with the same background data. Educational Testing Service's DGROUP program (Rogers et al., 2006) was applied with a two-dimensional model to generate the conditional distributions and PVs based on the context data and the

responses to the TIMSS & PIRLS linking items by fixing item parameters to the values from the operational TIMSS 2015 and PIRLS 2016, respectively.

For the ERCE math test, the posterior mean for each student was approximated by averaging the five PVs (the ERCE PVs were provided by ERCE team and on the ERCE reporting metric) from the ERCE mathematics scale. For the ERCE reading, the posterior mean for an individual was approximated by averaging the five PVs from the ERCE reading scale. The correlations between the posterior means of ERCE data and linking data (ERCE math and TIMSS linking, ERCE reading and PIRLS linking) range from 0.78 to 0.82. The concordance score range and levels were identified based on the estimated ERCE posterior means using the combined data of the two countries. The score ranges of the posterior means of the ERCE mathematics and reading scales were either rounded up or down to cover almost all the data of the two countries and to be as symmetric as possible around the overall mean of the ERCE scale (which is 700). For both ERCE scales, scores range from about 400 to 1000 (covering almost 100% of the data) with very few data points beyond the range of 440 to 940 (covering about 99.5% of the data). As a result, 26 score levels were identified in the score range of 440–940 for preliminary concordance table with 20 points apart for each of the two scales.

To construct a reliable concordance table, only students who participated in all four tests, ERCE math, ERCE reading, TIMSS linking, and PIRLS linking were included for the donor selection, *i.e*, 2619 students in Colombia and 3902 students in Guatemala. For each identified concordance score level, 5 donors were selected from each of the two countries so that each country contributes equally to the concordance tables. This selection was achieved by selecting the 5 smallest absolute differences of students' posterior mean on the ERCE test to each specified concordance score for each country. Each of the donors donated 5 PVs on the target tests. The mean and standard deviation of the donors' PVs from the Rosetta Stone linking data were calculated based on the total 50 donated PVs (2 countries * 5 donors * 5 PVs) at each level. These steps were implemented separately for ERCE mathematics and reading. In the following sections, concordance table for ERCE mathematics is used for illustration.

Preliminary concordance table for ERCE mathematics was created by assigning the estimated mean and standard deviation of each set of 50 PVs based on the Rosetta Stone linking data to each concordance level in the specified range of ERCE mathematics. The estimated mean and standard deviation of each set of PVs were weighted by using the total weight variable in the data and calculated with SAS version 9.4 (SAS Institute Inc). The mean and standard deviation of each set of 50 donated PVs were produced for each concordance score level between the range of 440 and 940 on the source test, ERCE math, as shown in Fig. 1. The projected conditional means based on the donated PVs on the target scale show that generally higher means are related to higher concordance scores for mathematics. Because of the volatility due to the limited number of countries and the smaller sample sizes available as donors per score point (not all students could be included in population modeling and donor selection), there was a small fluctuation at some concordance score points. Therefore, a smoothing procedure was used to better

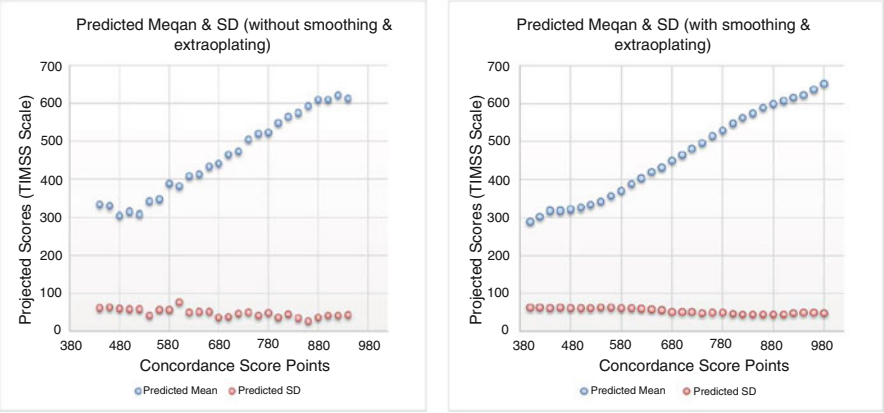


Fig. 1 Predicted Mean and SD with/without smoothing and extrapolating for ERCE math

represent the underlying projected conditional distributions on the target test. For each concordance score point, the mean of the donated PVs was smoothed by applying a simple moving average operation (*e.g.*, Isnanto, 2011) with a window of 7 score points. The smoothed mean X_i at a concordance score level, i, was calculated based on the unsmoothed mean x_i and the adjacent means as follows:

$$X_i = \frac{x_{i-3} + \cdots + x_i + \cdots + x_{i+3}}{7} \tag{7}$$

The standard deviation of PVs of each score point was smoothed in a similar way as the means of PVs as shown in Eq. (7) but with adjustments. First, getting the average of each set of the 7 variances (variance of the 50 donated PVs) clustered at the corresponding score level i in the concordance table for the corresponding variance,v_i. Next, the smoothed variance, V_i, was adjusted by adding the geometric mean of the smoothed variances for the all 26 score points (instead of arithmetic mean and was calculated by using GEOMEAN function in EXCEL) to better represent the variance of the PVs with the smaller sample sizes available as donors per score point. The square root of the final adjusted smoothed variance becomes the smoothed conditional standard deviation (SD) at that accordance score point.

To obtain a robust prediction for ERCE concordance scores beyond the range of 440–940, where only a very small number (less than 0.5%) of students was observed, a non-parametric regression method, Sen's slope estimator, was used to extrapolate for two more concordance score levels at both ends. To calculate the Sen's slope estimator for the predicted mean, the median of all slopes for all pairs of ordered ERCE score levels and the smoothed means were used to predict the conditional means of the likely posterior distributions at the concordance levels 400, 420, 960, and 980. The Sen's slope, for the ordered pairs (i, X_i) where X_i is the smoothed mean at the score level i, is calculated as:

Table 1 Concordance table for ERCE mathematics (partial table)

ERCE mathematics score	Projected score on TIMSS scale		Lower bound		Upper bound	
	Mean	SD	95%	68%	68%	95%
400	290	64	162	226	354	417
420	304	63	178	241	367	430
440	319	62	194	256	381	443
...						
680	449	53	344	397	502	555
700	465	52	362	414	517	569
720	481	51	379	430	532	583
...						
940	624	51	522	573	675	726
960	638	50	538	588	688	739
980	653	49	554	603	702	751

Note: More detailed information about Concordance Table for ERCE Mathematics and Reading can be found here: https://timssandpirls.bc.edu/Rosetta-Stone-Reports/index.html

$$Sen's\ slope = Median\left\{\frac{X_j - X_i}{j - i} : i < j\right\} \tag{8}$$

Similarly, the median of all slopes for all pairs of ordered score levels and the smoothed standard deviations were used to predict the conditional standard deviations. The Sen's slope estimators for the predicted mean and standard deviation are 14.47 and −0.85, respectively. Figure 1 shows the smoothed and unsmoothed predicted distributions for mathematics. The smoothed graph also included the four extrapolated score points at the two ends.

Table 1 shows part of the final concordance table for ERCE mathematics. The first column shows the ERCE concordance score levels. The second and third columns show the projected means and standard deviations (SDs) of the projected conditional distribution of the latent variable on the TIMSS scale given the ERCE score level. The last four columns show the lower and upper bounds (minimum and maximum values) of the 68% and 95% cut points for conditional distribution on TIMSS scale.

6 How to Use the Concordance Table

The concordance table can be used for estimating the percentages of students in each ERCE country reaching the four TIMSS 4th grade benchmarks. To estimate the percentages, a series of steps are needed to generate the projected PVs. First,

Table 2 Estimated percentages of 6th grade ERCE students reaching the 4th grade TIMSS international benchmarks

Country	Advanced (625)		High (550)		Intermediate (475)		Low (400)	
Estimated percentages based on Rosetta Stone								
Colombia	2.3	(0.5)	15.7	(1.4)	47.9	(2.4)	81.1	(1.7)
Guatemala	0.9	(0.3)	8.4	(0.9)	34.2	(1.6)	71.6	(1.7)
Average	**1.6**	**(0.3)**	**12.0**	**(0.8)**	**41.0**	**(1.4)**	**76.3**	**(1.2)**
Estimated percentages based on Concordance								
Colombia	2.7	(0.5)	16.4	(1.3)	48.2	(2.2)	81.3	(1.5)
Guatemala	1.1	(0.3)	8.1	(1.0)	30.4	(1.3)	66.6	(1.5)
Average	**1.9**	**(0.3)**	**12.3**	**(0.8)**	**39.3**	**(1.3)**	**73.9**	**(1.0)**

Note: Standard errors appear in parentheses

calculate the average of 5 PVs based on the ERCE sample in the domain of interest on the ERCE scale for each student. The average of the 5 PVs is the posterior mean of each student. Second, find the closest ERCE score level in the above concordance table for each student based on the estimated posterior mean. Third, assign the corresponding projected mean and SD on the TIMSS scale to each student based on the identified closest ERCE score level. Next, impute 5 new projected TIMSS PVs (target test) based on the assigned projected mean and SD for each student. PVs for individual students can be imputed using a normal distribution with the corresponding projected mean and SD. This step was repeated five times to get 5 PVs for each student. Then the percentages of students reaching the four TIMSS 4th grade benchmarks can be estimated based on the new projected 5 PVs.

The steps described here for generating projected PVs based on the concordance table and calculating the percentages of reaching the four benchmarks could be applied to all countries participated in ERCE 2019 but not participated in TIMSS assessment or Rosetta Stone study. This is one of the purposes for establishing the concordance table between two assessments. In this section, the percentages of 6th grade students participating in ERCE 2019 assessment reaching the four TIMSS 4th grade international benchmarks (Advanced: 625, High: 550, Intermediate: 475, Low: 400) were estimated and shown in Table 2 for two sets of PVs based on the same two countries and used to demonstrate the robustness of the concordance table. The first set of PVs are based on the Rosetta Stone linking data which were generated from population models, item responses together with the context data of students' background of information, before establishing the concordance table. The second set of PVs are the projected PVs based on the concordance table which were re-generated based on the steps described above.

Overall, Table 2 shows that while there is small variability in estimated countries' percentages when comparing the estimates based on concordance and those based on Rosetta Stone linking data, the average percentages across the two countries provide highly comparable results.

7 Conclusion and Recommendations

As large-scale assessments gain more popularity and publicity recently, interest in linking large-scale assessments from national, regional, to international has also grown. Several studies (Jia et al., 2014; Ehmke et al., 2020) attempted to link test scores from national assessments to international large-scale assessments using conventional equating-like linking methods. However, the construct equivalency assumptions associated with equating-like linking methods are typically unrealistic and not defensible in linking regional to international large-scale assessments because the tests being linked measure somewhat different constructs and are constructed in different ways.

This paper introduced a new approach for score projections by constructing an enhanced concordance table between two large-scale assessments with one source and one target tests. First, this new donor-based concordance approach integrates uncertainty due to construct differences and measurement error, and can be used to construct a concordance table for projections between two large-scale assessments. It appropriately provides a conditional distribution (with mean and SD) on the international target assessment given achievement on the regional source, rather than the point to point projection as in the conventional equipercentile linking methods. Second, the proposed concordance approach also avoids assumptions about statistical dependencies that rely on distributional assumptions or assumed complex functional forms as used in statistical moderations (Jia et al., 2014) to link the assessments. Most importantly, the constructed concordance table could also be applied to all countries participated in the same regional assessment but not participated in linking study due to financial reasons or other limitations for participation.

The Rosetta Stone ERCE linking study illustrates the robustness of the new approach by relating different regional assessment programs to TIMSS and PIRLS international long-standing metrics and benchmarks of achievement. The concordance tables enable educators and researchers to make inferences and interpret the ERCE results in relation to TIMSS and PIRLS international benchmarks of mathematical and reading knowledge comprehension.

However, the concordance should be used with care, being aware of the limitations of country participation and sample sizes, and differences between assessments. Also, concorded scores are not perfectly equivalent as they do not provide a direct link between assessments. It cannot be used as a point to point projection, either. Moreover, concordance tables vary by different samples as they are dependent on sample characteristics such as differences in school curricula, test language and language spoken at home, socioeconomic or sociodemographic differences. Even the uncertainty of the concordance has to be taken into consideration when constructing concordance tables. The concordance tables still may vary if sample characteristics in one country are very different from other participating countries. In addition, different choices for number of donors and intervals between concordance levels may also lead to differences in the projected

conditional distributions. Therefore, to link large-scale assessments in the future, educational practitioners and researchers are encouraged to use larger national sample sizes and add more countries in the linking study so that more donors are included from different countries to improve the estimated concordance and account for country-specific variability whenever it is possible. Further research on the impact of choices for number of donors and intervals between concordance levels on the resulting concordance is also warranted.

While the concordance has its limitations, it is an appropriate tool to allow comparisons between two large-scale assessments. It helps comparing difficulty levels between regional assessments and other large-scale assessments and allows studying the achievement distributions of them for benchmarking in educational monitoring.

References

Cartwright, F., Lalancette, D., Mussio, J., & Xing, D. (2003). Linking provincial student assessments with national and international assessments. In *Education, skills and learning, research papers* (Bd. 005). Statistics Canada.

Ehmke, T., van den Ham, A.-K., Sälzer, C., Heine, J., & Prenzel, M. (2020). Measuring mathematics competence in international and national large-scale assessments: Linking PISA and the national educational panel study in Germany. *Studies in Educational Evaluation, 65*, 100847.

Hernández-Torrano, D., & Courtney, M. G. R. (2021). Modern international large-scale assessment in education: An integrative review and mapping of the literature. *Large-Scale Assessments in Education, 9*(1), 17.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Praeger.

IEA website: https://www.iea.nl/studies/additionalstudies/rosetta

Isnanto, R. R. (2011). Comparation on several smoothing methods in nonparametric regression. *Jurnal Sistem Komputer, 1*(1), 41–47.

Jia, Y., Phillips, G., Wise, L. L., Rahman, T., Xu, X., Wiley, C., & Diaz, T. E. (2014). *2011 NAEP-TIMSS linking study: Technical report on the linking methodologies and their evaluations* (NCES 2014-461). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

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

Linn, R. L., McLaughlin, D., & Thissen, D. (2009). *Utility and validity of NAEP linking efforts*. American Institutes for Research, NAEP Validity Studies Panel.

Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics, 6*(3), 287–296. https://doi.org/10.2307/1391878

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

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. ETS Policy Information Center.

Nissen, A., Ehmke, T., Koller, O., & Duchhardt, C. (2015). Comparing apples with oranges? An approach to link TIMSS and the National Educational Panel Study in Germany via equipercentile and IRT methods. *Studies in Educational Evaluation, 47*, 58–67.

Rogers, A., Tang, C., Lin, M. J., & Kandathil, M. (2006). *DGROUP [Computer software]*. Educational Testing Service.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics, 4*(1), 87–94.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's Tau. *Journal of the American Statistical Association, 63*(324), 1379–1389.

UNESCO website: https://tcg.uis.unesco.org/rosetta-stone/

von Davier, M., & Sinharay, S. (2013). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment*. CRC Press.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? In M. von Davier & D. Hastedt (Eds.), *IERI monograph series: Issues and methodologies in large scale assessments* (Vol. 2, pp. 9–36). Hamburg.

Comparing Parametric and Nonparametric Methods for Heterogeneous Treatment Effects

Jee-Seon Kim, Xiangyi Liao, and Wen Wei Loh

Abstract Efforts to estimate treatment effects and draw causal inferences based on observational data are increasingly relevant with the abundance of such data in the social and behavioral sciences. Although the average treatment effect (ATE) might be the first step in the analysis, the main goal often concerns conditional average treatment effects (CATEs) of particular subgroups or treatment effects conditioning on a (set of) covariate(s). This study examines several parametric and nonparametric methods for CATE estimation. Specifically, we apply two machine learning methods, causal forest (CF) and Bayesian additive regression trees (BART), and two doubly-robust multilevel modeling approaches to the synthetic data used for the data challenge at the 2018 Atlantic Causal Inference Conference. We conclude with a discussion on the issues and challenges of different methods in estimating and interpreting CATE.

Keywords Conditional average treatment effects · Multilevel models · Hierarchical linear modeling · Machine learning · Propensity scores · Observational studies · Causal forest · Bayesian Additive Regression Trees (BART) · Clustered data · Doubly-robust estimators

1 Potential Outcomes and ATE in Clustered Data

Treatment effects can be defined using the Neyman-Rubin potential outcomes framework (Neyman, 1923; Rubin, 1974). We are using the extended notation of potential outcomes for the multilevel structure where units are nested within clusters (Hong & Raudenbush, 2006; Lyu et al., 2022). Assume that there are N individuals

J.-S. Kim (✉) · X. Liao
Department of Educational Psychology, University of Wisconsin, Madison, WI, USA
e-mail: jeeseonkim@wisc.edu; xliao36@wisc.edu

W. W. Loh
Department of Quantitative Theory and Methods, Emory University, Atlanta, GA, USA
e-mail: wen.wei.loh@emory.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_3

nested within M clusters. Let $Y_{ij}(1)$ denote the potential outcome if individual i within cluster j was treated ($T_{ij} = 1$) and $Y_{ij}(0)$ denote the potential outcome if individual i within cluster j was untreated ($T_{ij} = 0$), where $i = 1, \ldots, n_j$ in cluster $j = 1, \ldots, M$ and $\sum_{j=1}^{M} n_j = N$. The observed outcome can be presented as

$$Y_{ij} = T_{ij}Y_{ij}(1) + (1 - T_{ij})Y_{ij}(0)$$

under the *stable unit treatment value assumption* (SUTVA; Rubin, 1986) where the potential outcomes of each individual are not affected by others' treatment assignments, and there is only a single version of treatment. For more details on SUTVA for multilevel settings, see Hong and Raudenbush (2006, 2013).

As the two potential outcomes $Y_{ij}(0)$ and $Y_{ij}(1)$ are never observed simultaneously, individual treatment effects cannot be estimated. However, if the pair of potential outcomes $(Y_{ij}(0), Y_{ij}(1))$ is independent of treatment assignment T_{ij}, we can estimate average treatment effects (ATEs), the average linear contrast between two potential outcomes as:

$$\tau = E[Y_{ij}(1) - Y_{ij}(0)].$$

Block randomized experiments or multisite randomized trials achieve this independence through the randomization of treatment assignment within blocks or sites. For observational data, an *unconfoundedness* assumption is required to obtain the ATE, which implies that the potential outcomes are independent of treatment assignment, given the observed vector of covariates. Unconfoundedness is also referred to as *strong ignorability* in the causal inference literature (Rosenbaum & Rubin, 1983; Rubin, 1978), which implies that the potential outcomes are independent of treatment assignment, given observed individual covariates $\boldsymbol{X}_{ij}$ and cluster covariates $\boldsymbol{Z}_j$;

$$\text{Unconfoundedness} : Y_{ij}(1), Y_{ij}(0) \perp T_{ij} | \boldsymbol{X}_{ij}, \boldsymbol{Z}_j,$$

and the probability of each individual being assigned to a particular treatment given observed covariates is strictly between 0 and 1;

$$\text{Positivity or Overlap} : 0 < e(\boldsymbol{X}_{ij}, \boldsymbol{Z}_j) = \Pr(T_{ij} = 1 | \boldsymbol{X}_{ij}, \boldsymbol{Z}_j) < 1,$$

where $\perp$ denotes independence between two random variables and $e(\boldsymbol{X}_{ij}, \boldsymbol{Z}_j)$ is the *propensity score* (Rosenbaum & Rubin, 1983). Propensity scores are commonly estimated by logistic regression with single-level data and by random or fixed effects logistic regression with multilevel data (Leite, 2016; Fuentes et al., 2021). Propensity scores are often used in matching methods to match treated and control units or to weigh cases via inverse probability weighting towards eliminating confounding due to observed covariates and satisfying the unconfoundedness assumption (Kainz et al., 2017; Leite, 2016; Stuart, 2010).

2 Subgroup Analysis and CATE

When it is of interest to estimate subgroup-specific treatment effects that may differ from other subgroups or the ATE for the whole population, conditional Average treatment effects (CATEs) (Imbens & Rubin, 2015) can be estimated conditional on observed level-1 (e.g., individual) and level-2 (e.g., cluster) covariates, $\boldsymbol{X}_{ij}$ and $\boldsymbol{Z}_j$, respectively.

$$\tau_{ij} = \mathrm{E}[Y_{ij}(1) - Y_{ij}(0)|\boldsymbol{X}_{ij} = \boldsymbol{x}_{ij}, \boldsymbol{Z}_j = \boldsymbol{z}_j].$$

Machine learning methods, such as Bayesian additive regression trees (BART) and causal forests (CF), provide estimators of τ_{ij} under strong ignorability and SUTVA for observed covariates (Chipman et al., 2010; Wager & Athey, 2018). Multilevel models can also be used for the estimation of CATE, often in combination with propensity score adjustments such as matching, stratification, and weighting (Fuentes et al., 2021; Leite, 2016).

In this chapter, we consider CATE based on only observed covariates. The estimand cannot be estimated directly by multilevel models, CF, or BART alone when the subgroup memberships are unobservable or the relevant covariates are unmeasured. For the estimation and evaluation of treatment effects with unobserved subgroups, we refer readers to Kim and Steiner (2015), Kim et al. (2016), Suk et al. (2021), Lyu et al. (2022), and Loh and Kim (2022b).

3 Doubly-Robust Estimators Using Multilevel Models

3.1 Inverse Propensity Weighted Regression

Multilevel models have been used widely for analyzing nested or clustered data (Raudenbush & Bryk, 2002; Snijders & Bosker, 2011). To estimate CATE using multilevel models, we first fit fixed or random effects logistic regression to estimate propensity score $e(\boldsymbol{X}_{ij}, \boldsymbol{Z}_j)$ and then combine the use of propensity scores and the multilevel regression outcome. Such "doubly-robust" estimators are appealing because they offer protection against model misspecification biases when either model is correctly specified (Robins, 2000; Bang & Robins, 2005).

Let $\mu(T_{ij}, \boldsymbol{X}_{ij}, \boldsymbol{Z}_j) = \mathrm{E}[Y_{ij}|T_{ij}, \boldsymbol{X}_{ij}, \boldsymbol{Z}_j]$ denote the outcome function, the estimator of the CATE is then:

$$\widehat{\tau}_{ij}^{IPWReg} = \widehat{\mu}(T_{ij} = 1, \boldsymbol{X}_{ij}, \boldsymbol{Z}_j) - \widehat{\mu}(T_{ij} = 0, \boldsymbol{X}_{ij}, \boldsymbol{Z}_j). \quad (1)$$

We estimate μ using weighted multilevel regression with treatment indicator, all individual- and cluster-level covariates as well as (correctly specified) interaction terms. In particular, we weight each observation by the inverse propensity score

weight W_{ij}, which is defined as:

$$W_{ij} = \frac{T_{ij}}{e(\boldsymbol{X}_{ij}, \boldsymbol{Z}_j)} + \frac{1 - T_{ij}}{1 - e(\boldsymbol{X}_{ij}, \boldsymbol{Z}_j)}. \tag{2}$$

These weights are used to weigh the observed data when fitting the multilevel regression; see, e.g., Vansteelandt and Keiding (2011) for the single-level setting. The mean potential outcomes predicted using the fitted model are then used to construct the quantities in Eq. (1).

3.2 Augmented Inverse Propensity of Weighting

Loh and Kim (2022a), following Kang and Schafer (2007), explain three doubly robust estimation methods where the estimators differ in terms of how the estimated propensity scores are leveraged in the outcome model. Among the three methods, we adopt the *augmented inverse propensity weighted (AIPW)* estimator (Robins et al., 1994). The AIPW method imputes potential outcomes and avoids an explicit parametrization by augmenting the inverse probability weights estimator with an outcome model to exploit information about the treatment effects.

While the ATE is often encoded as the treatment coefficient in a parametric model for the outcome, the CATE is not as easily encoded as an explicit parameter in the outcome model. Furthermore, the AIPW estimator of the ATE is doubly robust in the sense that it is consistent if either the propensity score model or outcome model is correctly specified and asymptotically unbiased when both are correctly specified (Glynn & Quinn, 2010; Kurz, 2022). Therefore, the AIPW method is a particularly appealing doubly robust estimator for CATE.

The procedure of obtaining the CATE estimate using the AIPW estimator can be explained as follows: Let $\mu_1(\boldsymbol{X}_{ij}, \boldsymbol{Z}_j) = \mathrm{E}[Y_{ij}|T_{ij} = 1, \boldsymbol{X}_{ij}, \boldsymbol{Z}_j]$ and $\mu_0(\boldsymbol{X}_{ij}, \boldsymbol{Z}_j) = \mathrm{E}[Y_{ij}|T_{ij} = 0, \boldsymbol{X}_{ij}, \boldsymbol{Z}_j]$ be the outcome functions for treated and control units, respectively. The estimator of the CATE, among individuals with the same observed covariate values $(\boldsymbol{X}_{ij}, \boldsymbol{Z}_j)$, is then:

$$\begin{aligned} \widehat{\tau}_{ij}^{AIPW} =& T_{ij}\widehat{W}_{ij}(Y_{ij} - \widehat{\mu}_1(\boldsymbol{X}_{ij}, \boldsymbol{Z}_j)) + \widehat{\mu}_1(\boldsymbol{X}_{ij}, \boldsymbol{Z}_j) \\ & - (1 - T_{ij})\widehat{W}_{ij}(Y_{ij} - \widehat{\mu}_0(\boldsymbol{X}_{ij}, \boldsymbol{Z}_j)) - \widehat{\mu}_0(\boldsymbol{X}_{ij}, \boldsymbol{Z}_j). \end{aligned} \tag{3}$$

The estimates $\widehat{W}_{ij}$ are obtained by substituting the unknown propensity scores $e(\boldsymbol{X}_{ij}, \boldsymbol{Z}_j)$ in (2) with their estimates using, for example, a multilevel logistic regression with all individual- and cluster-level covariates. The estimates $\widehat{\mu}_1$ and $\widehat{\mu}_0$ are obtained using multilevel regression with all individual- and cluster-level covariates. In sum, the AIPW estimator lessens the reliance on traditional parametric models while profiting from correctly specifying either the selection model or the outcome model, but not necessarily both, for valid inference.

4 Data Challenge at a Causal Inference Conference

We revisit a workshop conducted at the 2018 Atlantic Causal Inference Conference (ACIC), where eight groups of researchers were invited to analyze synthetic data to assess treatment effect variation on an outcome (Carvalho et al., 2019). The generated dataset was motivated by the National Study of Learning Mindsets, a large-scale randomized trial of an online growth mindset intervention (Yeager et al., 2019).

Among several questions the participants were asked to address, we focus on the effect of the intervention in relation to a school-level variable we call "FIXED.MINDSET" (X1 in the paper), which is a measure of the average fixed mindset rating for each school before intervention. Specifically, we are interested in whether (1) the mindset intervention was effective in improving student achievement and (2) FIXED.MINDSET (school-level average fixed mindset score) moderates the effect of the intervention. It was found that while the ATE was very similar across the eight teams of researchers, CATE was substantially different depending on the approaches and methods used.

The organizer later revealed that the data were generated from the following model:

$$y_{ij} = \mu(\boldsymbol{x}_{ij}, \boldsymbol{z}_j) + \left[\tau_{ij} + U_{1j}\right] T_{ij} + U_{0j} + \epsilon_{ij},$$

where y_{ij} is the achievement score for student i in school j, μ is an additive function of student- and school-level covariates, U_{0j} and U_{1j} are random school effects that follow $N(0, 0.15^2)$ and $N(0, 0.105^2)$, respectively. Note that U_{0j} and U_{1j} were generated independently. Level-1 random effect ϵ_{ij} was drawn by jitter standard deviation 0.5. Treatment effects were generated as follows:

$$\begin{aligned}\tau_{ij} &= 0.228 + 0.05 \cdot \mathbb{1}(\text{FIXED.MINDEST} < 0.07) \\ &\quad - 0.05 \cdot \mathbb{1}(\text{ACAD.ACH} < -0.69) - 0.08 \cdot \mathbb{1}(\text{ETHNICITY} \in \{1, 13, 14\}),\end{aligned}$$

where ACAD.ACH is school achievement average before intervention and ETHNICITY is a categorical race/ethnicity variable. Therefore, for the two questions above, the correct ATE and CATE should reflect that (1) the mindset intervention improved student achievement and (2) the treatment was more effective (by 0.05 points in the outcome) for schools with fixed mindset scores lower (0.07 or less) at pretest. For further details of the data generation, see Carvalho et al. (2019).

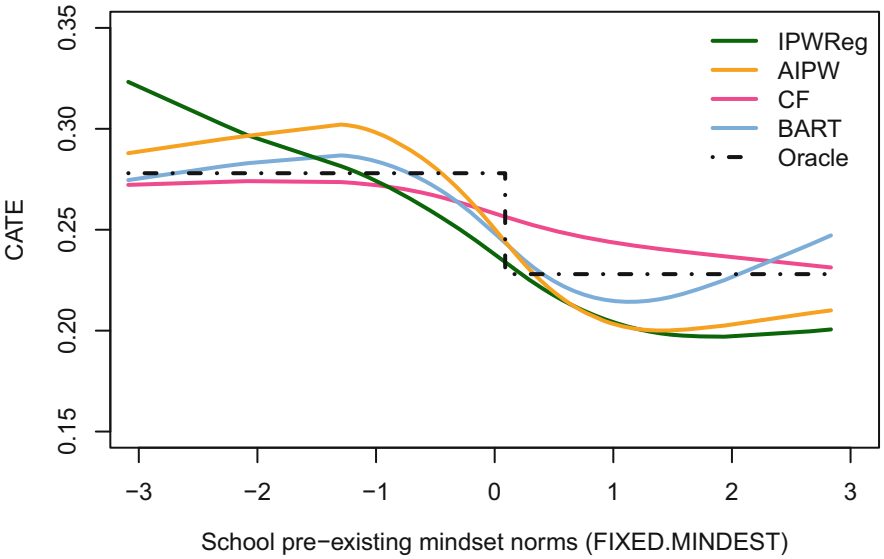


Fig. 1 Students' CATE estimates against school-level mindset. The dash-dotted line ("Oracle") represents the true CATE based on the data generating model. *IPWReg* Inverse Propensity Weighted Regression, *AIPW* Augmented Inverse Propensity of Weighting, *CF* Causal Forest, *BART* Bayesian Additive Regression Trees

5 CF, BART, and Multilevel Models for CATE

We reproduced the results of CF and BART as presented at the conference and subsequent dissemination (Athey & Wager, 2019; Carnegie et al., 2019), obtained two results using inverse propensity weighted regression (IPWReg) and the augmented inverse propensity of weighting (AIPW), and compared the four estimates to the "true" CATE based on the data generating model. All methods return similar ATE estimates around 2.5 and the results for CATE are depicted in Fig. 1.

Although none of the parametric and nonparametric methods reproduced the true step function at the change point of 0.07 perfectly, all four approaches discovered the decreasing trend of students' CATEs as a function of the school-level covariate and detected a sizable decline of the treatment effects around the school's average fixed mindset score of zero. We found that the CATE estimates using different methods deviated from the oracle values based on the data generating model in various ways. Specifically, CF closely resembled the range of the CATE but was insensitive to the sharp reduction in the middle. BART created a shape of a cubic function that was relatively close to the true CATEs for the low fixed mindset scores but not for the high scores. IPWReg and AIPW approximated students' CATEs around the inflection points closely but overestimated and underestimated the effects at the

lower and upper ends, respectively. IPWReg in particular showed the problem of extrapolation at the lower end, even after trimming extreme weights downward.

6 Discussion

Valid causal evaluations of conditional treatment effects based on observational studies require not only a procedure of controlling for potential confounding, but correctly specifying how the effect is modified (or moderated) by the covariates. Although numerous methods and procedures have been proposed to estimate CATEs, both parametric and nonparametric methods were built under strong assumptions and face different challenges. Parametric methods require correct specification of the model in addition to modeling assumptions. Nonparametric methods are generally more robust against model misspecification but harder to interpret than parametric models, and machine learning methods can be susceptible to overfitting.

We have compared different point estimators of the CATE. In future work, we will explore the statistical (and computational) efficiency of these estimators, by empirically investigating the coverage and widths of the corresponding confidence intervals (CIs) in finite samples. In particular, we will compare CIs constructed using either normal approximations with influence functions (for methods where these are available), or the nonparametric bootstrap; see e.g., Smith et al. (2022) for a comparison under the single-level setting.

With these noticeable differences between parametric and nonparametric methods in mind, this chapter compared the CATE estimates by four different methods; CF, BART, and two multilevel regression approaches. In our empirical comparison of these methods, we used the synthetic data created by the organizer of the data challenge workshop at the 2018 ACIC (Carvalho et al., 2019).

The data challenge at ACIC and our revisit of the workshop with additional methods highlight important issues in assessing heterogeneous treatment effects in non-randomized studies and even in randomized trials. As covariates are related to each other, whether a particular variable is a moderator of the treatment effect depends on whether the analysis conditions on related covariates or not, and it also depends on whether treatment effect variation was estimated across sampled clusters or in the population (Carvalho et al., 2019).

Examining heterogeneous treatment effects becomes more challenging in observational studies with an increasing number of covariates, due to the curse of dimensionality. Covariates may affect outcome and selection procedures differently and confounding may involve complex nonlinear relations, among many other reasons. Recent studies emphasize that it is crucial to consider multiple aspects of the methods and procedures such as sample splitting, inclusion of confounder interactions, and doubly robust estimators to obtain valid causal effect estimates and draw a proper inference, especially for subgroup effects and CATE (Naimi et al., 2021; Ratkovic, 2021).

References

Athey, S., & Wager, S. (2019). Estimating treatment effects with causal forests: An application. *Observational Studies, 5*(2), 37–51.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics, 61*(4), 962–973.

Carnegie, N., Dorie, V., & Hill, J. L. (2019). Examining treatment effect heterogeneity using BART. *Observational Studies, 5*(2), 52–70.

Carvalho, C., Feller, A., Murray, J., Woody, S., & Yeager, D. (2019). Assessing treatment effect variation in observational studies: Results from a data challenge. *Observational Studies, 5*(2), 21–35.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics, 4*(1), 266–298.

Fuentes, A., Lüdtke, O., & Robitzsch, A. (2021). Causal inference with multilevel data: A comparison of different propensity score weighting approaches. *Multivariate Behavioral Research, 57*(6), 916–939.

Glynn, A. N., & Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis, 18*(1), 36–56.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association, 101*(475), 901–910.

Hong, G., & Raudenbush, S. W. (2013). Heterogeneous agents, social interactions, and causal inference. In *Handbook of causal analysis for social research* (pp. 331–352). Springer.

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

Kainz, K., Greifer, N., Givens, A., Swietek, K., Lombardi, B. M., Zietz, S., & Kohn, J. L. (2017). Improving causal inference: Recommendations for covariate selection and balance in propensity score methods. *Journal of the Society for Social Work and Research, 8*(2), 279–303.

Kang, J. D., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science, 22*(4), 523–539.

Kim, J.-S., & Steiner, P. M. (2015). Multilevel propensity score methods for estimating causal effects: A latent class modeling strategy. In *Quantitative psychology research* (pp. 293–306). Springer.

Kim, J.-S., Lim, W.-C., & Steiner, P. M. (2016). Causal inference with observational multilevel data: Investigating selection and outcome heterogeneity. In *The Annual Meeting of the Psychometric Society* (pp. 287–308). Springer.

Kurz, C. F. (2022). Augmented inverse probability weighting and the double robustness property. *Medical Decision Making, 42*(2), 156–167.

Leite, W. (2016). *Practical propensity score methods using R*. Sage Publications.

Loh, W. W., & Kim, J.-S. (2022a). Causal models. In *International encyclopedia of education* (4th ed.). Elsevier.

Loh, W. W., & Kim, J.-S. (2022b). Evaluating sensitivity to classification uncertainty in subgroup effect analyses. *BMC Medical Research Methodology, 22*(1), 247.

Lyu, W., Kim, J.-S., & Suk, Y. (2022). Estimating heterogeneous treatment effects within latent class multilevel models: A bayesian approach. *Journal of Educational and Behavioral Statistics, 48*(1), 3–36.

Naimi, A. I., Mishler, A. E., & Kennedy, E. H. (2021). Challenges in obtaining valid causal effect estimates with machine learning algorithms. *American Journal of Epidemiology*. https://doi.org/10.1093/aje/kwab201.

Neyman, J. S. (1923). On the application of probability theory to agricultural experiments. essay on principles, section 9. *Statistical Science, 5*(4), 472–480

Ratkovic, M. (2021). Subgroup analysis: Pitfalls, promise, and honesty. *Advances in Experimental Political Science* (pp. 271–288). Cambridge University Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American statistical association* (Vol. 1999, pp. 6–10). Indianapolis.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association, 89*(427), 846–866.

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

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

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics, 6*(1), 34–58.

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association, 81*(396), 961–962.

Smith, M. J., Mansournia, M. A., Maringe, C., Zivich, P. N., Cole, S. R., Leyrat, C., Belot, A., Rachet, B., & Luque-Fernandez, M. A. (2022). Introduction to computational causal inference using reproducible stata, r, and python code: A tutorial. *Statistics in Medicine, 41*(2), 407–432.

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1–21.

Suk, Y., Kim, J.-S., & Kang, H. (2021). Hybridizing machine learning methods and finite mixture models for estimating heterogeneous treatment effects in latent classes. *Journal of Educational and Behavioral Statistics, 46*(3), 323–347.

Vansteelandt, S., & Keiding, N. (2011). Invited commentary: G-Computation–lost in translation? *American Journal of Epidemiology, 173*(7), 739–742.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association, 113*(523), 1228–1242.

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., et al. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature, 573*(7774), 364–369.

A Historical Perspective on Polytomous Unfolding Models

Ye Yuan and George Engelhard

Abstract This study provides a review and discussion of unfolding models for unidimensional polytomous data. Unfolding models (ideal point models) have a single-peaked response function. Unfolding models offer an underutilized and alternative approach to cumulative item response theory models for examining measurement data. Engelhard and Yuan (J Appl Measur, in press) described the basic principles of several key unfolding models for dichotomous responses. This study extends this work to graded responses. The study revisits main polytomous unfolding models including the Generalized Hyperbolic Cosine model (Andrich, Br J Math Stat Psychol 49(2):347–365, 1996), Graded Unfolding Model (Roberts and Laughlin, Appl Psychol Measur 20(3):231–255, 1996), Generalized Graded Unfolding Model (Roberts et al., Appl Psychol Measur 24(1):3–32, 2000), and nonparametric unfolding models for multicategory data (van Schuur, Polit Anal 4:41–74, 1992). One of the major goals of this study is to highlight the underlying principles, formulations, measurement properties, and implementations of selected polytomous unfolding models. The main purpose of this study is to call attention to the use of unfolding models for polytomous responses for modeling measurement data. The study also highlights the importance of using cumulative versus unfolding models for attitude measurement.

Keywords Unfolding model · Attitude measurement · IRT model

The basic underlying idea of unfolding models can be traced back to Thurstone and Chave (1929). Thurstone distinguished between maximum probability and increasing probability scales. Increasing probability scales undergird various cumulative item response models including the Rasch model and other IRT models. The maximum probability scales are called a variety of names including unfolding, ideal-point, non-monotonic, non-cumulative and proximity scales. One of the

Y. Yuan (✉) · G. Engelhard
Department of Educational Psychology, University of Georgia, Athens, GA, USA
e-mail: ye.yuan@uga.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_4

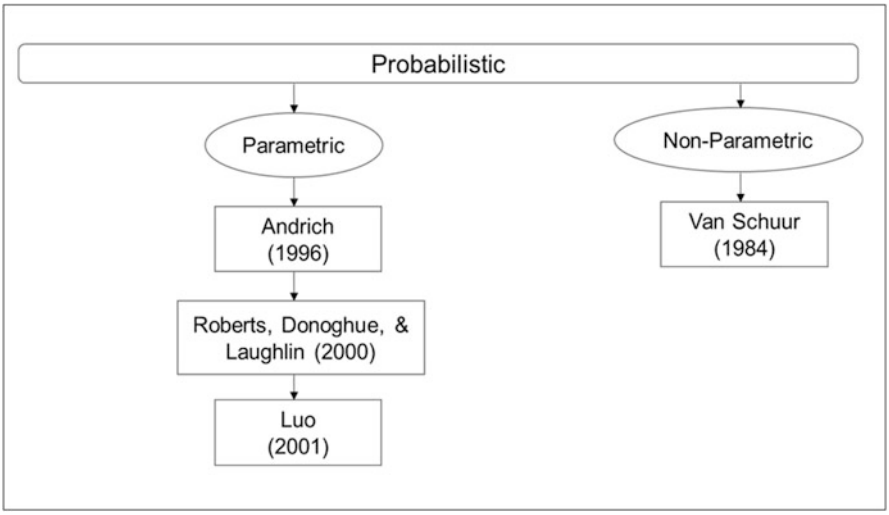


Fig. 1 Polytomous unfolding models taxonomy

fundamental features of an unfolding response process is that the probability of a positive response is a single-peaked function (Coombs & Avrunin, 1977). Unfolding models have been used in a variety of areas with a major focus on attitude measurement.

Significant research has discussed the structure of polytomous cumulative IRT models, such as nominal categories item response model (Bock, 1972), the Samejima's graded response model (Samejima, 1969), the Rasch rating scale model (RSM; Andrich, 1978; Rasch, 1961), and the generalized partial credit model (GPCM; Muraki, 1992). However, unfolding models with maximum probability scales have received less attention. It is worth revisiting the polytomous unfolding models in order to foster interest in using unfolding models.

The purpose of study is to provide a brief historical perspective on polytomous unfolding models. Unfolding models are an underutilized approach for modeling response data. A main theme of this study is to call attention to the importance of unfolding models for polytomous responses as an alternative to cumulative models.

This study discusses unidimensional unfolding models that all have a fundamental characteristic of a unimodal response process. Figure 1 shows a basic taxonomy of unfolding models for polytomous data. Engelhard and Yuan (in press) have summarized a taxonomy of dichotomous unfolding models that shows the distinctions between deterministic and probabilistic models, nonparametric and parametric models. Figure 1 highlights the development and extension of dichotomous responses models into polytomous rating. Based on this taxonomy, the present study describes important polytomous unfolding models and compares them on several characteristics.

1 Description of Polytomous Unfolding IRT Models

GHCM The Hyperbolic cosine model (HCM; Andrich & Luo, 1993) is an unfolding model for dichotomous responses. Andrich (1996) proposed a general hyperbolic cosine model for unfolding polytomous responses (GHCM). The model is a generalized HCM designed to reconcile Thurstone and Likert approaches. Under the GHCM for unfolding Likert-style responses, the distance between thresholds should be symmetrical. For example, Fig. 2 (Panel A) shows a case with 4 observed categories and 7 latent response categories. It is unknown if the person responds to the strongly disagree/disagree below or above the statement location, thus the distance of thresholds, such as $\tau_2 - \tau_1$ should equal to $\tau_6 - \tau_5$. The key idea of this generalization is to fold over the latent categories based on the equations shown in Fig. 2 (Panel B). It is analogous to the dichotomous case, but the equation can be extended to ordered response categories to calculate the probabilities of disagree by summing up the "disagree below" and "disagree above". The equations in Panel B are based on the Rasch model for ordered response categories (Andrich, 1978), where β_n is the person location and δ_i is the statement location on the continuum. The parameter $k_{yi} = -\sum_{k=1}^{y} \tau_{ki}$ represents the thresholds. A normalizing constant γ_{ni} is used to keep $\sum_{y=0}^{2m} P\{Y_{ni} = y\} = 1$, where $m + 1$ represents the number of observed response categories. The construction can be similarly generalized as the form of GHCM as:

$$P\{X_{ni} = x; x < m\} = \frac{1}{\gamma_{ni}} (\exp k_{xi})\ 2\cosh[(m - x)(\beta_n - \delta_i)],$$

$$P\{X_{ni} = m\} = \frac{1}{\gamma_{ni}} (\exp k_{mi})$$

Figure 3 (Panel A) presents an example of the GHCM's item category response curve.

GUM. The graded unfolding model (GUM; Roberts & Laughlin, 1996) is another unfolding form for polytomous responses. The difference from GHCM is that in the context of GUM, the most positive response $X_{ni} = m$ (such as "strongly agree") is also considered as two possible responses. The key idea of GUM is based on the rating scale model (Andrich, 1978) and the probabilities of two subjective responses:

$$P(X_i = x|\beta_n) = P(Y_i = x|\beta_n) + P[Y_i = (2m + 1 - x)|\beta_n]$$

The equations yield the form of GUM:

$$P(X_{ni} = x) = \frac{\exp\left[x(\beta_n - \delta_i) - \sum_{k=0}^{x} \tau_k\right] + \exp\left[(M - x)(\beta_n - \delta_i) - \sum_{k=0}^{x} \tau_k\right]}{\sum_{w=0}^{m} \left\{\exp\left[w(\beta_n - \delta_i) - \sum_{k=0}^{w} \tau_k\right] + \exp\left[(M - w)\left((\beta_n - \delta_i) - \sum_{k=0}^{w} \tau_k\right]\right\}}$$

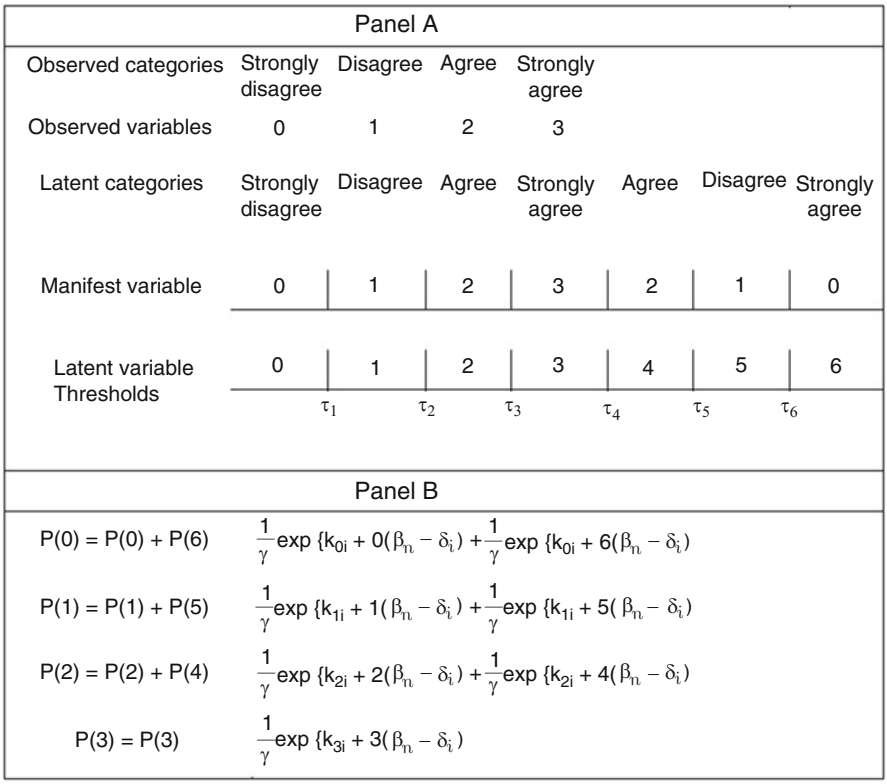


Fig. 2 Illustrative of GHCM for polytomous responses

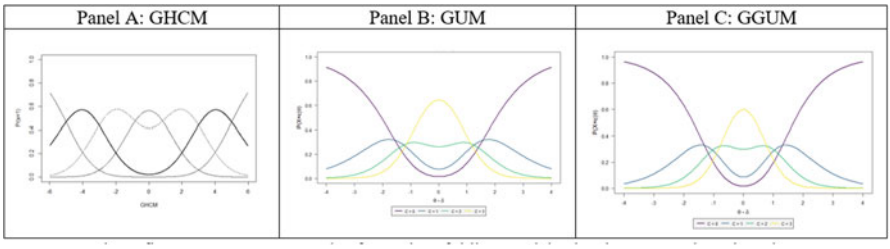


Fig. 3 Illustrative of item category response functions
Note. These figures represent examples for each unfolding model. The shapes vary based on the values of the parameters in the model, and these figures were selected to illustrate distinctive shapes for each model

where $M = 2m + 1$. Figure 3 Panel B presents a GUM item category response curve using the GGUM R package. Luo (2001) re-expressed GUM from another perspective. Luo (2001) explained the rating formulation approach and the process of mapping a polytomous response that we will discuss later. A general form of polytomous unfolding models was generated in Luo (2001), where ψ is the

operational function. Engelhard and Yuan summarized the operational functions for dichotomous unfolding models (Engelhard & Yuan, in press). Luo found that the GUM equation is a case of this general form of polytomous unfolding models. Interested readers can consult Luo (2001) for details of the probabilistic function and the operational function satisfying the properties (Luo, 2001).

GGUM The Generalized graded unfolding model (GGUM; Roberts et al., 2000) is a generalization of the GUM. The GGUM family of models are flexible in analyzing dichotomous and polytomous unfolding responses data. Like GUM, GGUM also considers each observable response category correspond to two unique subjective response categories. The subjective response categories are the latent categories which follow a cumulative IRT model, for examples, a rating scale model or partial credit model. Figure 3 (Panel C) shows the item category response curve of GGUM. GGUM adds an item discrimination parameter (α_i) to the model based on the generalized partial credit model (Muraki, 1992) for the subjective response categories. When θ_j is the location of the *jth* individual on the latent continuum, the probability of a person endorsing a particular observable response category is the sum of the probabilities of the two subjective responses. The definition of GGUM can be expressed as,

$$P\left(Z_i = z|\theta_j\right) = \frac{\exp\left\{\alpha_i\left[z\left(\theta_j - \delta_i\right) - \sum_{k=0}^{z}\tau_{ik}\right]\right\} + \exp\left\{\alpha_i\left[(M-z)\left(\theta_j - \delta_i\right) - \sum_{k=0}^{z}\tau_{ik}\right]\right\}}{\sum_{w=0}^{C}\left\{\exp\left\{\alpha_i\left[w\left(\theta_j - \delta_i\right) - \sum_{k=0}^{w}\tau_{ik}\right]\right\} + \exp\left\{\alpha_i\left[(M-w)\left(\theta_j - \delta_i\right) - \sum_{k=0}^{w}\tau_{ik}\right]\right\}\right\}}$$

where Z_i is an observable response to statement i, M represents the number of the latent response categories minus 1, and C is the number of observable response categories minus 1 (Roberts et al., 2000). Marginal maximum likelihood can be used to estimate the item parameters for GGUM. An expected a posteriori procedure can be used to estimate the person parameters.

Luo's General Form Luo (2001) constructed a general class of polytomous unfolding models. We showed its approach to GUM in the above section. Luo (2001) proposed a general formulation of polytomous unfolding models. Using this general form, a Hyperbolic cosine model for polytomous responses (HCM-P model; Luo, 2001), a polytomous unfolding models for simple squared logistic model (SSLMP; Luo, 2001), and a PARELLA model for polytomous responses can be developed. The definitions of these models are summarized in Table 1.

Nonparametric Method The aforementioned models use parametric methods. This section introduces a nonparametric unfolding model and its extension for graded responses data. Multiple unidimensional unfolding (MUDFOLD; van Schuur, 1984) is a nonparametric method to analyze dichotomous data. Nonparametric methods can select items and identify a homogenous set of indicators (van Schuur, 1992). The nonparametric approach does not explicitly formulate a measurement model. In the cumulative data, the Mokken model (Mokken, 1971) can be viewed as

Table 1 Luo's general form of polytomous unfolding models

Dichotomous	**HCM (Andrich & Luo, 1993)**	**SSLM (Andrich, 1988)**	**PARELLA (Hoijtink, 1990)**
	$P\left(Z_{nik}=1\right)=\frac{\cosh(\rho_{ik})}{\cosh(\rho_{ik})+\cosh(\beta_n-\delta_i)}$, $k=1,\ \ldots,m.$	$P\left(Z_{nik}=1\right)=\frac{\exp(\rho_{ik}^2)}{\exp(\rho_{ik}^2)+\exp\left[\left((\beta_n-\delta_i)^2\right.\right.}$, $k=1,\ \ldots,m.$	$P\left(Z_{nik}=1\right)=\frac{{\rho_{ik}}^2}{{\rho_{ik}}^2+(\rho_n-\delta_i)^2}$, $k=1,\ \ldots,m.$
Polytomous	**HCM-P (Luo, 2001)**	**SSLMP (Luo, 2001)**	**PARELLAP (Luo, 2001)**
	$P\left(X_{ni}=k\right)=$ $\frac{[\cosh(\beta_n-\delta)]^{m-k}\prod_{l=1}^{k}\cosh(\rho_{il})}{\lambda_{ni}}$, $k=0,\ \ldots,m-1.$	$P\left(X_{ni}=k\right)=$ $\frac{\exp\left\{\sum_{l=1}^{k}\rho_{il}^2\right\}\exp\left\{(m-k)(\beta_n-\delta_i)^2\right\}}{\lambda_{ni}}$, $k=0,\ \ldots,m-1.$	$P\left(X_{ni}=k\right)=\frac{(\beta_n-\delta)^{2(m-k)}\prod_{l}^{k}{\rho_{ik}}^2}{\lambda_{ni}}$, $k=0,\ \ldots,m-1.$
	$\lambda_{ni}=$ $\sum_{k=0}^{m}\left[\cosh\left(\beta_n-\delta\right)\right]^{m-k}\prod_{l=1}^{k}\cosh\left(\rho_{il}\right)$	$\lambda_{ni}=$ $\sum_{k=0}^{m}\exp\left\{\sum_{l=1}^{k}\rho_{il}^2\right\}\exp\left\{(m-k)\left(\beta_n-\delta_i\right)^2\right\}$	$\lambda_{ni}=\sum_{k=0}^{m}\left(\beta_n-\delta\right)^{2(m-k)}\prod_{l=1}^{k}{\rho_{ik}}^2$
	Luo (2001)		
General Form of Polytomous Unfolding Models	$P\left\{X_{ni}=k\middle\vert\beta_n,\delta_i,(\rho_{il})\right\}=$ $\frac{\left(\prod_{l=1}^{k}\psi_l(\rho_{il})\right)\left(\prod_{l=k+1}^{m}\psi_l(\beta_n-\delta_i)\right)}{\lambda_{ni}}$, $k=0,\ldots,m$ $\lambda_{ni}\equiv$ $\sum_{k=0}^{m}\left(\prod_{l=1}^{k}\psi_l\left(\rho_{il}\right)\right)\left(\prod_{l=k+1}^{m}\psi_l\left(\beta_n-\delta_i\right)\right)$		

Note. The parameters are defined in the test

a nonparametric method and the Rasch model is a parametric counterpart . For unfolding response processes, the MUDFOLD approach assumes that the item characteristics curves are single-peaked and provides the order of the items along the latent trait. Besides the common assumptions of unidimensionality and local independence, the unfolding IRT model has the assumption of a unimodal function of theta for every item. The stochastic ordering and the manifest unimodality are two other assumptions. Van Schuur (1992) extended the MUDFOLD for multicategory data. This extension is analogous to Molenaar's (1982) extendsion of Mokken's (1971) nonparametric unidimensional cumulative model. The procedure for conducting MUDFOLD for multicategory data is similar to MUDFOLD for dichotomous data. Corresponding to Guttman's model for the triangular pattern, the response patterns (1,0) are not permitted. For examples, (1,0,1,1) has one error, and (1,0,1,0,1) has two errors (Leik & Matthews, 1968). In the context of multicategory data, there are different error response patterns that violate the unfolding model. For example, the response categories are 0, 1, and 2, pattern 202 indicates four error patterns: 202, 201, 102, and 212. The total number of errors in each response pattern can be calculated by summing up the number of errors in each triple of items over all triples (van Schuur, 1992). The calculation of the expected number of errors in each triple of items and the ordering of items in an unfolding scale for multicategory data are similar to the calculation for the dichotomous data. The examination of goodness of fit including the assessment of H-coefficients and the probability pattern. A positive H-coefficient value is expected. The probabilities for the highest or the lowest response are expected to follow the same characteristic monotonicity pattern as in the dichotomous cases. Also, the interpretations of the dominance matrix, the adjacency matrix, the conditional adjacency matrix, and the correlation matrix are the same as for dichotomous data (van Schuur, 1992).

2 Discussion

This study briefly describes the main unfolding models for unidimensional polytomous responses. Polytomous unfolding models, such as GGUM, have been used in attitude and personality assessments for years. Other models also have great potential to contribute to both cognitive and non-cognitive assessments. Future research on polytomous unfolding IRT models can pay more attention to implementing various models to support and improve education and social science measurement from the psychometric perspective.

References

Andrich, D. (1988). The application of an unfolding model of the PIRT type to the measurement of attitude. *Applied Psychological Measurement, 12*(1), 33–51.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–573.

Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology, 49*(2), 347–365.

Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement, 17*(3), 253–276.

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

Coombs, C. H., & Avrunin, G. S. (1977). Single-peaked functions and the theory of preference. *Psychological review, 84*(2), 216.

Engelhard, G., & Yuan, Y. (in press). A historical perspective on unfolding models. *Journal of Applied Measurement*.

Hoijtink, H. (1990). PARELLA: Measurement of latent traits by proximity items. The Netherlands: University of Groningen, 1990.

Leik, R. K., & Matthews, M. (1968). A scale for developmental processes. *American Sociological Review, 33*, 62–75.

Luo, G. (2001). A class of probabilistic unfolding models for polytomous responses. *Journal of Mathematical Psychology, 45*(2), 224–248.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. De Gruyter Mouton.

Molenaar, I. W. (1982). Mokken scaling revisited. *Kwantitatieve methoden, 3*(8), 145–164.

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

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (pp. 321–334). University of California Press.

Roberts, J. S., & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement, 20*(3), 231–255.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*(1), 3–32.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, Monograph, 17*, 34.

Thurstone, L. L., & Chave, E. (1929). *The measurement of attitudes*. University of Chicago Press.

Van Schuur, W. H. (1984). *Structure in political beliefs: A new model for stochastic unfolding with application to European party activists*. CT Press.

Van Schuur, W. H. (1992). Nonparametric unidimensional unfolding for multicategory data. *Political Analysis, 4*, 41–74.

Kernel Equating Presmoothing Methods: An Empirical Study with Mixed-Format Test Forms

Joakim Wallmark, Maria Josefsson, and Marie Wiberg

Abstract When equating test forms, it is common to presmooth the test score distributions before conducting the equating. In this study, the log-linear and item response theory (IRT) presmoothing methods were compared when equating mixed-format test forms using kernel equating. Test forms from two different high-stakes tests were equated: The Swedish national test in mathematics, using the equivalent group sampling design, and the verbal part of the Swedish SAT test, using the nonequivalent groups with anchor test sampling design. In both cases, the analytical equating standard errors were lower for high and low performing test takers when using IRT presmoothing compared to log-linear presmoothing. Both presmoothing methods resulted in reasonable equated curves. As no true equating transformation is known in a practical setting, using IRT models for presmoothing appears to be a viable alternative to log-linear models when equating mixed-format tests such as the Swedish SAT.

Keywords Kernel equating · Presmoothing · Item response theory

1 Introduction

Large-scale and high-stakes testing programs typically require construction of multiple forms of the same test. It is common to compare test forms from different administrations using *test score equating* (Kolen & Brennan, 2014; González & Wiberg, 2017). Test forms can be constructed in different ways using different types of items, potentially affecting which equating methods can be used and their efficiency. For example, the items on each test form may be multiple choice items, scored dichotomously, or constructed response items, scored polytomously. In general multiple-choice items require shorter testing time, while constructed response items can measure a deeper level of understanding and reasoning. It has

J. Wallmark (✉) · M. Josefsson · M. Wiberg
Department of Statistics, USBE, Umeå University, Umeå, Sweden
e-mail: joakim.wallmark@umu.se; maria.josefsson@umu.se; marie.wiberg@umu.se

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_5

also been shown that multiple-choice items may provide less information about high and low performing test takers when compared to constructed response items (Ercikan et al., 1998). Since time is often limited and tests aim to cover a broad range of knowledge in a topic, a mix of different item types are sometimes used in the same test form. Such test forms are commonly referred to as mixed-format tests (Ercikan et al., 1998; Kim et al., 2008, 2010a,b; Kolen & Lee, 2014). The National Assessment of Educational Progress, the Advanced Placement Program, the SAT Reasoning Test and the national test in mathematics in Sweden are all examples of mixed-format tests. Despite the popularity of mixed-format tests, earlier research on mixed-format test equating has mostly considered traditional equating methods (see e.g. Kolen & Lee, 2014). Kernel equating (von Davier et al., 2004b) has increased in popularity in recent years. The method is flexible and has been shown to perform well for both small and large sample sizes, making it an attractive alternative to traditional- (e.g. Kolen & Brennan, 2014) and item response theory (IRT; Lord, 1980) equating methods.

As a first step when using kernel equating, the test score distributions are typically smoothed out to reduce the impact from sampling error on the equated scores. This procedure is commonly referred to as *presmoothing*. Historically, the most common way to presmooth the data has been through the use of log-linear models. A lot of research on kernel equating has been conducted using this method of presmoothing, which we will refer to as log-linear kernel equating (LLKE) (e.g. Mao et al., 2006; von Davier et al., 2006; Moses et al., 2007; Liu and Low, 2008). More recently, Andersson and Wiberg (2017) proposed the use of IRT models for presmoothing of test forms containing dichotomously scored items. The same IRT presmoothing method was extended to include polytomous items using polytomous IRT models (Andersson, 2016). Limited research has been conducted to evaluate the performance of kernel equating with IRT presmoothing when polytomous IRT models are used on real test data.

The aim of this study is to evaluate the performance of kernel equating using IRT presmoothing on real test data. The resulting equating transformations were compared against log-linear presmoothing alternatives. Test data from the Swedish national test in mathematics as well as the Swedish scholastic aptitude test (SAT) were equated. The Swedish national test was equated using the equivalent groups (EG) design, in which the test taker samples taking each test form are assumed to be sampled from the same population. The SAT was equated using the nonequivalent groups with anchor test (NEAT) design. Under the NEAT design, population differences are adjusted for using a set of common items, typically referred to as the anchor test, given to both test taker groups.

In Sect. 2, descriptions of kernel equating as well as IRT- and log-linear presmoothing are given. This is followed by an empirical study, with methodology and results in Sects. 3 and 4. Finally, advantages, limitations and practical implications of the two presmoothing methods are discussed in Sect. 5.

2 Kernel Equating

Let X and Y denote two test forms, administered to samples from population P and Q respectively. Under the NEAT design, let A denote the anchor test form administered to both samples. Further, we denote the discrete cumulative distribution functions (cdfs) for the test scores on X and Y by $F_X(x)$ and $F_Y(y)$. For the anchor test, let $F_{AP}(a)$ and $F_{AQ}(a)$ denote the cdfs for the test scores on A in population P and Q respectively.

The goal in equating is to equate the scores on X to Y using the equipercentile transformation $\varphi(x) = F_Y^{-1}(F_X(x))$ for a target population T. This transformation only exists for continuous cdfs $F_X(x)$ and $F_Y(y)$. Kernel equating (von Davier et al., 2004b) uses kernel smoothing to estimate $\varphi(x)$ using continuous approximations of the typically discrete test score distributions in the samples. Different *kernels* can be used in this process, the most common being the Gaussian kernel (von Davier et al., 2004b; Mao et al., 2006; Moses et al., 2007; von Davier et al., 2006). The continuous approximation of $F_X(x)$ when using a Gaussian kernel is

$$F_{h_X}(x) = \sum_{j=0}^{K} r_j \Phi\left(\frac{x - \delta_X x_j - (1-\delta_X)\mu_X}{\delta_X h_X}\right), \tag{1}$$

where $\Phi(\cdot)$ is the standard normal distribution function and K is the total score on form X (assuming only non-negative integers are possible X scores). x_j is the jth score value, r_j is the probability for the jth score value, μ_X is the mean of the X scores, h_X is the bandwidth and $\delta_X = \sqrt{\frac{\sigma_X^2}{\sigma_X^2 + h_X^2}}$ where σ_X^2 is the variance of the form X scores. The approximations of $F_Y(y)$, $F_{AP}(a)$ and $F_{AQ}(a)$ are obtained in the same fashion, and denoted $F_{h_Y}(y)$, $F_{h_{AP}}(a)$ and $F_{h_{AQ}}(a)$. The bandwidth parameter h_X determines the smoothness of $F_{h_X}(x)$ and can be selected by minimizing the function

$$\text{PEN}(h_X) = \sum_{j=0}^{K}\left[r_j - \frac{d}{dx}F_{h_X}(x_j)\right]^2. \tag{2}$$

When using kernel equating under the NEAT design, one can use either the chained equating (CE) method or the Post-stratification equating (PSE) method. When using PSE, the anchor test scores are used to compute $F_{h_X}(x)$ and $F_{h_Y}(y)$ directly for a weighted target population $T = wP + (1-w)Q$ where $0 \le w \le 1$ is the weight given to P. When using CE, the equating transformation is obtained by chaining separate equatings $\hat{\varphi}(x) = F_{h_Y}^{-1}(F_{h_{AQ}}(F_{h_{AP}}^{-1}(F_{h_X}(x))))$. See von Davier et al. (2004a) for additional details.

Before approximating the continuous distribution functions, the discrete score distributions are typically *presmoothed* to reduce the random errors due to sampling. The log-linear and IRT model presmoothing methods are described below.

2.1 Log-Linear Presmoothing

When using log-linear models under the EG design, the frequencies of each X score n_x are modelled by $\log(n_x) = \beta_0 + \sum_{l=1}^{D_1} \beta_l x^l$ where D_1 is the highest polynomial degree. Under the NEAT design, let n_{xa} be the frequency of a score combination with x and a. The model can now be written $\log(n_{xa}) = \beta_0 + \sum_{l=1}^{D_1} \beta_l^X x^l + \sum_{l=1}^{D_2} \beta_l^A a^l + \sum_{l=1}^{D_3} \sum_{m=1}^{D_4} \beta_{lm}^{XA} x^l a^m$ where D_1, D_2, D_3 and D_4 denote the maximum polynomial degrees. The X, A, and XA superscripts are used to distinguish between the model parameters β associated with the powers of the X scores, the A scores, and the cross products of the X and A scores, respectively.

2.2 IRT Presmoothing

Any IRT model that can be fit to the data can be used in the presmoothing step. When dealing with polytomous data, a common option is the generalized partial credit (GPC) model (Muraki, 1992), defined as

$$P_{im}(\theta) = \begin{cases} \dfrac{1}{1 + \sum_{g=1}^{M_i - 1} \exp\left(\sum_{t=1}^{g} \left[a_i \left(\theta - b_{it}\right)\right]\right)}, & \text{if } m = 1 \\ \dfrac{\exp\left(\sum_{t=1}^{m} \left[a_i \left(\theta - b_{it}\right)\right]\right)}{1 + \sum_{g=1}^{M_i - 1} \exp\left(\sum_{t=1}^{g} \left[a_i \left(\theta - b_{it}\right)\right]\right)}, & \text{otherwise} \end{cases} \tag{3}$$

where i denotes the item, M_i is the number of response categories, $P_{im}(\theta)$ is the probability that a test taker with ability θ responds in response category m, b_{it} is the item category difficulty parameter and a_i is the item discrimination parameter. Note that this model generalizes to the two parameter logistic model for items with only two response categories (Lord, 1980). When using the GPC model, θ is assumed to be uni-dimensional and the responses to different items are assumed independent conditional on θ. After fitting the model to each item, the total score probabilities conditional on θ can be computed using the algorithm introduced by Thissen et al. (1995). As a final step, the marginal total score probabilities, r_j in Eq. (1), can be retrieved by integrating out the latent trait θ, thus averaging over the population.

3 Empirical Study

To compare IRT with log-linear presmoothing, test forms from the Swedish national test in mathematics (NAT) as well as forms from the Swedish SAT were equated. The national mathematics test is given to high-school students taking the mathematics 3c course. The test has a large impact on the course grade, and

Table 1 Summary statistics for the equated test forms

	EG		NEAT			
Statistic	X_{NAT}	Y_{NAT}	X_{SAT}	Y_{SAT}	A_{SAT}	A_{SAT}
Year	2019	2018	2014	2013	2014	2013
Total score	58	57	80	80	40	40
Number of items	28	28	60	60	30	30
Dichotomous items	9	9	50	50	25	25
Polytomous items	19	19	10	10	5	5
Mean	25.54	25.72	40.12	40.94	17.56	17.71
Standard deviation	12.62	12.12	12.88	13.34	7.06	7.13
Anchor test correlation	–	–	0.85	0.86	–	–
Sample size	1401	1008	2859	2469	2859	2469

the grade is later used together with the grades of other courses to apply for university programs. It is a mixed-format test with different types of items, requiring either short answers or step-by-step solutions. Some items are polytomously scored while others are scored dichotomously. There are no anchor items available for the national mathematics test forms. As the populations of test takers are similar in age and from similar educational background we assumed that the populations were equal and the test form from 2019 (X_{NAT}) was equated to the 2018 form (Y_{NAT}) under the EG design.

Using a set of anchor items, the SAT test form from 2014 (X_{SAT}) was equated to the 2013 (Y_{SAT}) form under the NEAT design. Both the CE and PSE methods were compared. The Swedish SAT is given twice a year, and the scores are used for applying to university. At least one third of the spots on different educational programs are given to students taking the SAT, while the remainder apply using their high school grades. The SAT consists of a verbal part and a quantitative part, which are equated separately. In this study, only the verbal part was equated. Three different types of items are contained within the test form: sentence completion, word interpretation and reading comprehension. All items are multiple choice items. Before conducting IRTKE, the scores from the reading comprehension items referring to the same texts were added together to form polytomous items, as the responses on these items cannot be assumed to be independent.

Summary statistics for all test forms are displayed in Table 1. Among the polytomous items on X_{NAT} and Y_{NAT}, 12 items had three response categories. On X_{NAT}, there were three items with four response categories and four items with five response categories. However, Y_{NAT} had four four-category items and three five-category items, resulting in the total score on X_{NAT} being one point higher. On both Swedish SAT test forms, there were six items with three response categories, two items five categories and two items with six. As shown in Table 1, most items on the SAT forms were dichotomous.

The kequate (Andersson et al., 2013) R package was used to equate the test forms. GPC models were used for IRT presmoothing, see Eq. (3). For LLKE, order

four polynomial log-linear models were used to model the score frequencies on each test form under the EG design. The polynomial orders were chosen based on Akaike information criterion (AIC, Akaike, 1981). The Bayesian information criterion (BIC, Schwarz, 1978) has been shown to be more efficient than AIC for bivariate smoothing (Moses & Holland, 2010), and was used for polynomial order selection under the NEAT design. For both SAT forms, this resulted in degree five polynomials for the scores on the main test forms, degree four polynomials for anchor item scores, and cross products with maximum powers of two for the main form scores along with power one anchor scores.

Penalty minimization based on minimizing Eq. (2) was used for bandwidth selection. For method evaluation, the standard error of equating (SEE) of each equating transformation was compared together with the equated scores.

4 Results

For all equated test forms, both presmoothing methods result in somewhat similar amounts of smoothing. As an example, Fig. 1 shows the presmoothed score distributions from using IRT and log-linear presmoothing on X_{NAT} and Y_{NAT}. For these forms, the score frequencies in the data show large differences between adjacent score points, especially towards the middle part of the score distributions. The differences between IRT presmoothing and log-linear presmoothing appear to be relatively small. The largest difference is on the lower end of X_{NAT} score scale, where the IRT presmoothing method puts the frequency of a score of zero 5.78 points below the log-linear frequency.

Figure 2 displays each estimated equating transformation alongside its corresponding analytical SEE for the national mathematics test. Figure 3 shows the same for the Swedish SAT equatings. In the left plots in each figure, the score on X_{NAT}/X_{SAT} is subtracted from the equated scores to better visualize the differences between each equating method. A negative difference suggests lower score on Y_{NAT}/Y_{SAT} for a given score on X_{NAT}/X_{SAT}.

When looking at the NAT equating results in Fig. 2, one should note that the total score on X_{NAT} was one score point higher than the total score on Y_{NAT}. For this test, both IRT and log-linear presmoothing give an indication that the scores on X_{NAT} correspond to a marginally higher Y_{NAT} scores on the lower part of the score scale. The curves are similar in shape everywhere except for on the upper end, where the IRTKE equating transformation changes direction and has a downward slope around a X_{NAT} score of 40. Despite the overall similarities between the presmoothing methods in Fig. 1, the equated scores differ with more than 1.5 score points around an X score of 52. The reason being that multiple subsequent IRT score frequencies are below the corresponding log-linear frequencies around a score of 40 on X_{NAT}. This results in relatively large differences in the score distribution percentiles which persist throughout the kernel smoothing step. On average, the equating transformation obtained using LLKE was 0.23 score points

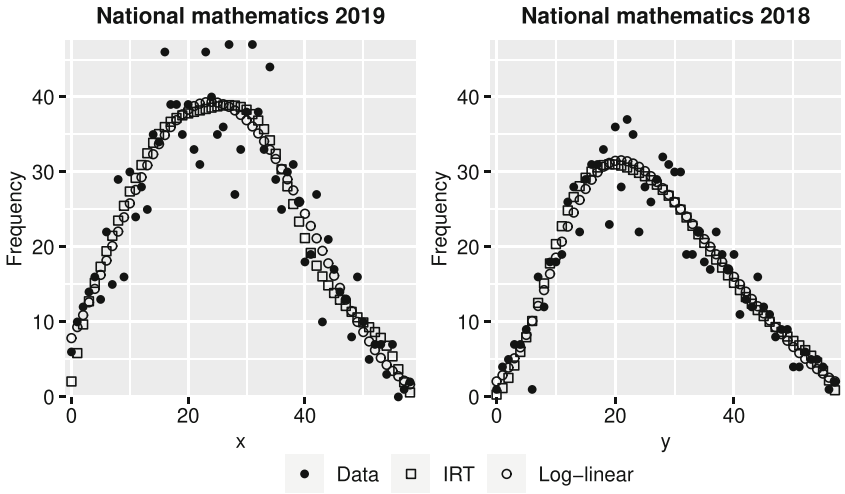


Fig. 1 The presmoothed score frequencies from each method are displayed together with the score frequencies in the data

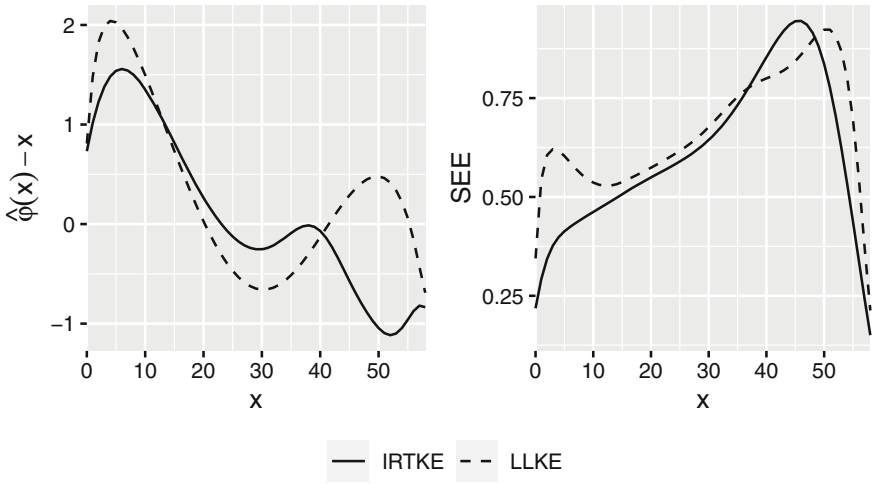


Fig. 2 The left plot shows the differences between the estimated equating transformations and their corresponding score x on the national mathematics test form from 2019. The plot to the right shows the SEEs associated with each curve

above the IRTKE transformation. One interpretation would be that using log-linear presmoothing makes Y_{NAT} appear easier compared with using IRT presmoothing for this test. As displayed in the right plot in Fig. 2, the estimated SEEs are lower for the majority for most of the score scale when using IRT presmoothing compared

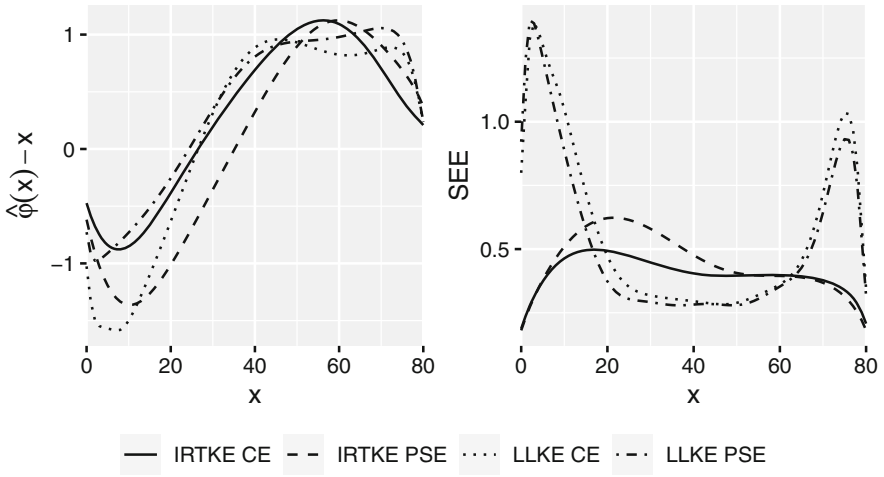


Fig. 3 The left plot shows the differences between the estimated equating transformations and their corresponding score x on the SAT test form from 2013. The plot to the right shows the SEEs associated with each curve

to using log-linear models. The largest differences are at the ends of the score scales. On the lower end, the SEEs from using IRT are almost half the size of the LLKE SEEs for some X_{NAT} scores. The average SEE was 0.61 for the IRTKE equating and 0.67 for the LLKE equating.

Under the NEAT design, Fig. 3, the differences between each curve is smaller even though the total test score is higher and more methods are compared. The equated scores are within one point of their corresponding scores on X_{SAT} over the whole score scale for all equating methods. The curves are also similar in shape and the largest differences are found towards the lower end of the score scale. All methods suggest that a score over 36 on X_{SAT} equates to a marginally higher score on Y_{SAT}. The resulting equating transformations from IRTTKE CE and LLKE PSE are relatively close to each other for most of the X_{SAT} score scale. These transformations differ mostly for the very best and the worst performing test takers. There appears to be no clear separation of the equating transformations based on presmoothing method or whether PSE or CE was used.

Looking at the SEEs in the right plot of Fig. 3, it is clear that for the lower and upper X_{SAT} scores, the SEEs from using IRT presmoothing are much larger than when using log-linear models. The opposite is true for the mid score range, but to a smaller extent. Over the entire score scale, the average SEEs were lower when presmoothing using IRT. The averaged SEEs were 0.40, 0.45, 0.57 and 0.52 for IRTKE with CE, IRTKE with PSE, LLKE with CE and LLKE with PSE respectively.

5 Discussion

In this study, kernel equating with IRT and log-linear presmoothing was compared when equating mixed-format tests using real test data under both the EG and NEAT data collection designs. Despite some differences, the equated scores from both presmoothing methods seemed plausible for all equated tests. There is no obvious reason as to why one method should be preferred over another just by looking at the equated scores. See for example Fig. 3.

The SEEs at the upper and lower ends of the score scales were smaller when using IRT presmoothing compared to log-linear presmoothing for both the SAT and the NAT. Andersson and Wiberg (2017) observed the same phenomenon in their simulation study using only dichotomous items, also comparing LLKE with IRTKE. This could possibly be explained by the fact that the log-linear method models the total test score distribution directly, without the need for specific items scores. With these models, having a few extra test takers at the top or bottom scores can result in "bumps" in the smoothed distributions, later impacting the equated scores. On the contrary, when using IRT presmoothing, the probabilities of responding in a certain response category for each item are modelled first, and forced to follow the parametric forms of the chosen IRT model. These smooth curves are then combined to get the smoothed total score distribution, imposing further smoothing at the lower and upper ends of the score scales, which in turn results in smaller SEEs. Whether this extra smoothing is desirable depends on whether or not the inconsistencies between the score frequencies at the top or bottom scores are there by chance or if it is a feature in the underlying population which should be modelled. However, as the true distribution is in practice unknown, the lower SEEs from IRT presmoothing may be attractive if equated scores of top performing students are of interest.

In this study, real test data was equated when comparing different methods. A limitation with this approach is that the true equating transformation is unknown, making it hard to judge which method performed better in terms of the equated scores. A simulation study would solve this issue, even though it is sometimes hard to come up with a fair true equating transformation in a simulated setting. Additionally, we only compared different equating methods within the kernel equating framework. In future studies, kernel equating should be further compared to other equating methods when test forms contain polytomous or mixed-format items to see if similar findings would be obtained. For example, Wang et al. (2020) showed that IRT observed-score equating produced better results compared to both IRTKE and LLKE in various scenarios under the NEAT design for tests containing only dichotomous items.

Another limitation is that only unidimensional GPC models were considered when equating using IRTKE. It would be of interest, especially in a mixed-format setting, when different items can sometimes be assumed to measure slightly different constructs, to explore the performance when using multidimensional IRT models. Additionally, in a mixed-format setting, the use of mixed IRT models could be considered (Chon et al., 2010). With this approach, a three parameter logistic

model could be used for dichotomous items to incorporate guessing, while the GPC or any other polytomous model could be used for the polytomous items. Non-parametric models (e.g. Wiberg et al., 2019; Tsutsumi et al., 2021) could also be used in situations where the parametric models do not fit the data very well.

In conclusion, using IRT models for presmoothing when using kernel equating on mixed-format test forms appears to be a viable alternative to using log-linear models. However, since no true equating is known, multiple equating methods should be considered and compared even if one method has lower SEEs. This could help judge if the resulting equating transformations appear plausible. For example, if multiple methods are compared and one method gives vastly different results while the other ones are more similar, it could be an indication that the deviating one is wrong, unless one has specific reasons to believe otherwise.

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

References

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of econometrics*, *16*, 3–14.

Andersson, B. (2016). Asymptotic standard errors of observed-score equating with polytomous IRT models. *Journal of Educational Measurement*, *53*(4), 459–477.

Andersson, B., & Wiberg, M. (2017). Item response theory observed-score kernel equating. *psychometrika*, *82*(1), 48–66.

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

Chon, K. H., Lee, W.-C., & Dunbar, S. B. (2010). A comparison of item fit statistics for mixed IRT models. *Journal of Educational Measurement*, *47*(3), 318–338.

Ercikan, K., Sehwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement*, *35*(2), 137–154.

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

Kim, S., Walker, M. E., & McHale, F. (2008). Equating of mixed-format tests in large-scale assessments. *ETS Research Report Series*, *2008*(1), i–26.

Kim, S., Walker, M. E., & McHale, F. (2010a). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement*, *47*(1), 36–53.

Kim, S., Walker, M. E., & McHale, F. (2010b). Investigating the effectiveness of equating designs for constructed-response tests in large-scale assessments. *Journal of Educational Measurement*, *47*(2), 186–201.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking*. Springer.

Kolen, M. J., & Lee, W.-C. (2014). *Mixed-format tests: Psychometric properties with a primary focus on equating*. CASMA Monograph No. 2.3 (Vol. 3). The University of Iowa.

Liu, J., & Low, A. C. (2008). A comparison of the kernel equating method with traditional equating methods using SAT® data. *Journal of Educational Measurement*, *45*(4), 309–323.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (Zeroth ed.). Routledge.

Mao, X., von Davier, A. A., & Rupp, S. (2006). Comparisons of the kernel equating method with the traditional equating methods on praxis™ data. *ETS Research Report Series*, *2006*(2), i–31.

Moses, T., & Holland, P. W. (2010). A comparison of statistical selection strategies for univariate and bivariate log-linear models. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 557–574.

Moses, T., Yang, W.-L., & Wilson, C. (2007). Using kernel equating to assess item order effects on test scores. *Journal of Educational Measurement*, *44*(2), 157–178.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, *1992*(1), i–30.

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

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, *19*(1), 39–49.

Tsutsumi, E., Kinoshita, R., & Ueno, M. (2021). Deep item response theory as a novel test theory based on deep learning. *Electronics*, *10*(9), 1020.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004a). The chain and post-stratification methods for observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement*, *41*(1), 15–32.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004b). *The kernel method of test equating*. Statistics for social science and public policy. Springer.

von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2006). An evaluation of the kernel equating method: A special study with pseudotests constructed from real test data. *ETS Research Report Series*, *2006*(1), i–31.

Wang, S., Zhang, M., & You, S. (2020). A comparison of IRT observed score kernel equating and several equating methods. *Frontiers in Psychology*, *11*, 308.

Wiberg, M., Ramsay, J. O., & Li, J. (2019). Optimal scores: An alternative to parametric item response theory and sum scores. *Psychometrika*, *84*(1), 310–322.

Equating Different Test Scores with Landmark Registration Compared to Equipercentile Equating

Marie Wiberg, James O. Ramsay, and Juan Li

Abstract The overall aim of this study is to propose a new test score equating method based on landmark registration, which has its roots in functional data analysis. A further aim is to compare sum scores and binary optimal scores and examine the proposed method in comparison with equipercentile equating with both simulated data and data from a college admissions test. We used the EG design in both the empirical study and the simulation study. In the simulation study, we examined the behaviour when either difficulty or discrimination differ in the test forms or if there is a large ability difference between the groups taking different test forms. The results indicate that the proposed method worked well. The empirical study suggested that the proposed method could be used in practice. Practical implications of using different test scores in test equating as well as the usefulness of landmark registration was discussed.

Keywords Landmark registration · Equipercentile equating · EG design

1 Introduction

There are several options when calculating test takers' test scores. If the used test has binary scored items, we could use sum scores, which is simple the sum of the correct answers. Advantages of sum scores include that they are computationally fast, easy to calculate and easily understood by the test takers and the general public. Another

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

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

J. Li
Ottawa Hospital Research Institute, Ottawa, ON, Canada
e-mail: juli@ohri.ca

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_6

option for calculating test scores are to use parametric item response theory (IRT) and calculate the test takers' latent test score. The advantage is that the test takers ability and the test scores are placed on the same scale but a large disadvantage is they are computationally demanding. A third option is to use optimal scores obtained from nonparametric item response functions (Ramsay & Wiberg, 2017a,b; Wiberg et al., 2018). Optimal scores are obtained by optimizing a criterion for fitting the data from a test taker within some class of nonparametric IRT models. Optimal scores have been shown to be more efficient than sum scores in terms of bias and root mean squared error (Ramsay & Wiberg, 2017b) and less computationally demanding than scores from parametric IRT (Wiberg et al., 2019). However, in order for optimal scores to be useful in testing programs we need to have tools to compare test scores between different test forms.

In test score equating we use statistical models and methods to make test scores comparable among different test forms so that the scores can be used interchangeably (González & Wiberg, 2017). A large number of equating methods have been developed for various data collection designs and different test scores. González and Wiberg (2017) divided these into traditional methods, such as mean, linear and equipercentile equating (Kolen & Brennan, 2014), kernel equating methods (von Davier et al., 2004) and parametric IRT methods (Lord, 1980). The most common method to use with binary scored items is equipercentile equating and this method can be used regardless of the test score used.

Another possibility if we want to place test scores on the same scale is by using the methodology from functional data analysis which optimal scores are built upon. In functional data analysis, so-called registration is typically used to put two functions on the same scale. Common registration methods include shift registration, curve registration, and landmark registration (Ramsay & Silverman, 2005). In this study we will use landmark registration. The overall aim of this study is to propose a new test score equating method based on landmark registration and compare it with the commonly used equipercentile test equating method using both real data and a simulation study in the equivalent groups (EG) design. A further aim is to compare sum scores and binary optimal scores and examine the proposed method in comparison with equipercentile equating. Examined conditions include when either difficulty or discrimination differ in the test forms and if there is a large ability difference between the groups taking different test forms.

Previous studies on optimal test scores have focused on developing the theory and comparisons with sum scores (Ramsay & Wiberg, 2017a,b), comparisons with parametric IRT (Wiberg et al., 2019) and developing the theory behind optimal scores further (Ramsay et al., 2020b). So far, no studies have been focused on equating optimal test scores. Previous studies on test score equating has mainly focused on the use of sum scores or IRT scores, for a summary see e.g. Kolen and Brennan (2014) and for implementations refer to González and Wiberg (2017).

The rest of the paper is structured as follows. In the next section sum scores and optimal scores are briefly described, followed by a section which contains a short description of equipercentile test equating. In Sect. 4, we propose to use landmark registration of curves in order to equate binary optimal scores or sum scores. Section 5 contains an empirical example with a real college admission

test where binary optimal scores and sum scores are compared when using both equipercentile equating and landmark registration in an EG design. The sixth section contains a simulation study and the seventh section displays the results from the simulation study. Finally, a discussion section with some concluding remarks and practical implications are given.

2 Sum Scores and Optimal Scores

For simplicity, we focus on test data which are binary scored, but our approach can be easily modified to optimal scores based on full information by taking account test takers wrong answers (see Ramsay et al. (2020b). Assume we have response data $U_{ij} = 0$ or 1, where $i = 1, \ldots, n$ are the items and $j = 1, \ldots, N$ the test takers. To calculate the sum score S_j for test taker j, we sum the response pattern over the i items, $S_j = \sum_{i=1}^{n} U_{ij}$.

Let θ denote a measure of a test taker's ability in terms of a test score. To calculate optimal scores, we use the negative log likelihood as fitting criterion for an arbitrary item response function over the interval $[0, n]$ defined as

$$-\log L(\theta|U) = \sum_{i=1}^{n}[-U_i W_i(\theta) + \log(1 + \exp W_i(\theta))], \tag{1}$$

where the item function $W_i(\theta)$ is the log-odds ratio

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

and $P_i(\theta)$ is the probability of answering an item correctly. To estimate $W_i(\theta)$, B-spline basis function expansions are used

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

where for each item i, γ_{ik} is the coefficient of the basis function ψ_{ik} in the basis function expansion of the ith item characteristic curve. To obtain optimal scores we use the first derivative of the negative log likelihood at the optimal θ defined as

$$\sum_{i=1}^{n} U_{ij}\frac{dW_i}{d\theta} - \sum_{i=1}^{n} P_i(\theta)\frac{dW_i}{d\theta} = 0. \tag{4}$$

The first two terms on the left hand side is a weighted sum of the data, and at the optimal θ the two terms are equal. The n weights $dW_i/d\theta$ are the slopes of the log-odds functions at θ, and decides the importance of each term in the sums. Thus, the quality of the information provided by a response is measured by how fast the

log-odds is increasing or decreasing at θ. For more details on how to obtain optimal scores refer to Ramsay and Wiberg (2017a) and for a comparison with sum scores refer to Wiberg et al. (2018) and for a comparison with parametric IRT scores refer to Wiberg et al. (2019). To obtain optimal scores in practice we use the R package TestGardener (Ramsay & Li, 2021).

3 Equipercentile Equating

Assume that we want to equate test form X to test form Y. Let X be the test scores from test form X and let Y be the test scores from test form Y. Let F_X and G_Y be the test scores cumulative distribution functions (CDF's) and F_X^{-1} and G_Y^{-1} their functional inverses, so that $F_X^{-1}[F_X(x)] = (F_X^{-1} \circ F_X)(x) = x$. Let $\varphi(x)$ denote the equating transformation, which for an equipercentile equating is defined as

$$\varphi(x) = G_Y^{-1}(F_X(x)) = (G_Y^{-1} \circ F_X)(x), \tag{5}$$

(Braun & Holland, 1982). The equating transformation will have different appearance depending on the test equating method and data collection design and thus there exist a large amount of methods to equate test scores.

In the EG design, we assume that the two samples who take test form X and test form Y are from the same population P. Scores from the different test forms at the same quintiles are assumed to be equivalent and thus we can use the equating transformation defined in (5) directly.

When using optimal scores with these traditional equating transformations we simply exchange the use of sum scores with optimal scores when we calculate the equating transformation.

4 Equating Test Scores with Landmark Registration

As discussed above, in functional data analysis registration is typically used to put two functions on the same scale. In this study we use landmark registration, which means that for each curve, argument values are identified which are associated with some features (Ramsay & Silverman, 2005). An advantage with landmark registration is its speed and that it allows for a continuous registration process. Landmark registration uses points to remove phase variation by transforming the domain of each curve so the location of shape features are aligned across curves. A typical choice of landmarks are percentiles such as the 5, 25, 50, 75 and 95 percentiles but one can place as many landmarks as one believes is needed. Landmark registration can be used as a fast low-dimensional approximation of the inverse in Eq. (1).

Let the CDF's of scores on two test forms be $F_X(x)$ and $G_Y(y)$, respectively. For simplicity we will assume that the test scores X and Y are over the same range, which we denote by $[0, \tau]$. Let function $Y = h(X)$ map test scores in the X-space into test scores in Y-space, where in this case the scales are the same, but they could be different. Let $\phi(X)$ be a strictly increasing smooth function of test scores, labelled *warping function*, such that $\phi(0) = 0$ and $\phi(\tau) = \tau$. Let the CDF F_X be warped into target CDF G_Y in the sense that

$$G_Y[\phi(x)] \approx F_X(x). \tag{6}$$

This warping function is similar to the equipercentile equating definition in (1). Landmark registration can thus be seen as a kind of equipercentile equating where the difference lies in how the equating transformation is obtained. The warping function in (6), is our new proposed equating transformation.

In the spirit of nonparametric analysis of test data, we need to represent $F_X(x)$ and $G_Y(y)$ as having an arbitrary level of accuracy. The accuracy will usually be defined by the test analyst by requiring a balance between how well (6) approaches an equality and how smooth the two CDF's and the warping function are. The two most important factors in this choice are likely to be the numbers of test takers and the numbers of items in the two test forms.

A CDF F and a warping function ϕ share three features. Each is required to be strictly increasing, each must be constrainable to be as smooth as is required, and each is normalized: $F(0) = \phi(0) = 0$, and $F(\tau) = 1$ while $\phi(\tau) = \tau$. We achieve these characteristics by the following transformations introduced by Ramsay (1996), which define them in terms of two unconstrained functions $L_F(x)$ and $T_\phi(x)$:

$$F(x) = \int_0^x \exp[L_F(u)]\,du \;\text{ and }\; \phi(x) = \tau \frac{\int_0^x \exp[T_\phi(u)]\,du}{\int_0^\tau \exp[T_\phi(u)]\,du}. \tag{7}$$

Each function is a strictly increasing function of the score x by virtue of the fact that the indefinite integral of a positive function will necessarily increase. In the case of the CDF, we see that function L_F is simply the log-density function $L_F(x) = \log dF/dx$, and this description is also reasonably appropriate for warping function ϕ since the larger $T_\phi(x)$ is the faster $\phi(x)$ increases.

The required flexibility in either the CDF or the warping function is achieved by defining the corresponding log-density function L as a basis function expansion

$$L(x) = \sum_k^K c_k \xi_k(x). \tag{8}$$

B-spline basis functions are the usual choice for the basis functions ξ_k, and the larger their number K, the more flexible the log-density. An explicit roughness penalty on the size of the second or higher derivatives of the spline expansion can also be used to impose further smoothness in the context of high-dimensional basis

expansions. A single-parameter warping function that is contained within this class is

$$\phi(x) = \begin{cases} \tau \frac{e^{vx}-1}{e^{v\tau}-1}, & \text{if } v \neq 0, \\ x, & \text{if } v = 0. \end{cases}$$

The restriction $F(0) = \phi(0) = 0$ is automatically satisfied by the definition in (7). For the estimation of (8), the fitting criterion is the negative log likelihood, the coefficients are optimized subject to the linear restriction $\sum_k c_k = 0$ and the normalizing constant is computed on each iteration. In the warping case, the optimization is subject to the restriction $v = 0$. Further details and illustrations are provided by Ramsay and Silverman (2005) and Ramsay et al. (2009) and functions coded for these purposes are available in the fda package available for both R and Matlab and can be obtained from the website www.functionaldata.org. In the EG design, the described warping function can be used directly.

5 Empirical Study

To illustrate the proposed equating method and in order to compare the use of sum scores and optimal scores in the EG design we examined real test data from the Swedish scholastic aptitude test (SweSAT) from two consecutive administrations (labelled 13B and 14A) referred to as the test forms. SweSAT is a binary-scored multiple-choice test that are used for college admissions. The test is typically given twice a year and it consists of a verbal and a quantitative subtest with 80 items each. We used the quantitative part with an EG design, and used two samples of 2000 test takers, who had taken either of two test forms. To items are the same on the test forms but the test forms are built to have similar items. We equated the new test form 14A to the old test form 13B.

Both sum scores and binary optimal scores were used. The test scores were used with the traditional equipercentile equating method and with the proposed equating method based on landmark registrations of the CDFs. We used a large number of basis functions, i.e. 15 to approximate the empirical CDF's.

Everything was done in R with the R packages *TestGardener* (Ramsay & Li, 2021) to estimate optimal scores, *equate* (Albano, 2016) to conduct equipercentile equating and *fda* (Ramsay et al., 2022) to obtain the warping functions. The used code can be obtained from the authors upon request.

5.1 Empirical Study Results

Figure 1 displays the warping function (i.e. the equating transformation) for landmark registration when test form 14A sum scores are warped to the score scale of the 13B test form. Note, the diagonal line is the identity function. A general

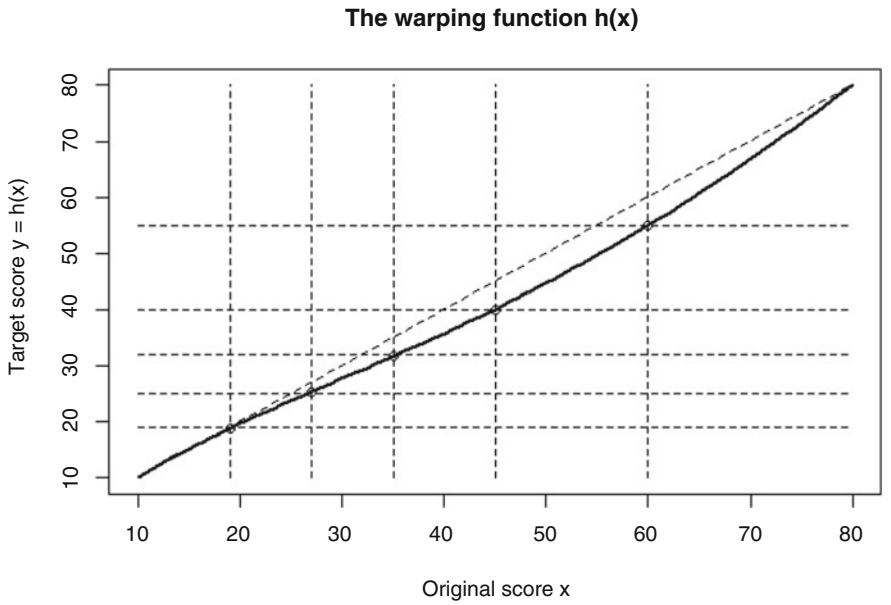


Fig. 1 The warping function for landmark registration from original 14A sum scores to the target 13B sum scores. The vertical lines represent the 5, 25, 50, 75 and 95 percentiles

results is that the 14A test form gives higher test scores than the 13B test form over the whole scale. Notice that the smallest test score was 10 rather than 0 because examinees were encouraged to guess if they couldn't make a choice otherwise. Most items had four response options, except 12 items which had five response options. Even if an examinee was guessing at all 80 questions, it would be highly improbable that the test score would be close to zero. Guessing is a form of contamination of the choice data that misrepresents the performance level of the weakest examinees. The warping function for optimal scores had a similar appearance and was thus excluded here.

Note that the warping function is concave because it shifts the original 14A score values to the right. The registration shows that a 14A test taker scoring 60 would score 55 on the 13B test, and a 14A test taker scoring 45 would score about 40 on the 13B test. The warping penalizes the 14A examinees for its easier items. Interestingly to note, the equipercentile equating illustrates the opposite pattern. A test taker would have received higher scores on 13B than on 14A. The reason is that the test scores functions are built differently.

In order to examine the differences more closely, refer to Table 1, which gives the equated values at each of the examined percentiles for all the examined methods. Reasonable equated scores were obtained when optimal scores were used with the equipercentile equating methods. Not surprisingly, the equated values are most affected by the choice of test scores and less effective of the choice of equating

Table 1 Selected percentiles (Perc), original test form scores (14A) and equated values for sum scores and optimal scores using landmark registration (LR) and equipercentile (EP) equating with the EG design

EG design	Sum scores			Optimal scores		
Perc	14*A*	*LR*	*EP*	14*A*	*LR*	*EP*
0.05	19	19	18	9	9	6
0.25	27	25	24	21	24	21
0.50	35	32	31	35	38	36
0.75	45	40	39	48	53	50
0.95	60	55	55	61	64	62

method. Using optimal scores instead of sum scores seems to give in general similar test scores in the lower score range (first quartile) but larger test score differences at the end points. Noticeable is that the proposed method yielded reasonable equated scores regardless of test score in the lower score range where we only had a few test takers. The results of using landmark registration to equate the test scores was similar to the results from equipercentile equating with optimal scores for the boundary values but differed somewhat in the mid score range. Not shown here as we only showed results up to the 95 percentile, the proposed method forces the highest score to be no larger than the highest score 80.

6 Simulation Study

To illustrate how the equating is affected by different test scores and to compare the proposed method with equipercentile equating for the different test scores we used a simulation study. In the simulation study we assume that we equate from test form X to test form Y and we simulated binary response data for a 80 items multiple choice test. We sampled 5000 test takers for each test forms X and Y. We used a baseline case, which we obtained by simulating item response data with a two-parameter logistic IRT model and the ability of the test takers were drawn from N(0,1). We choose to use landmarks at the following percentiles; 0.05, 0.25, 0.50, 0.75, 0.95. Note, one can easily choose to use more landmarks if there are certain percentiles which are of higher interest, such as grading limits. The five landmarks chosen here were only used to illustrate that a large number of landmarks is not needed to get stable equating results.

The discrimination parameters were drawn from a U(0.3,1.3) distribution, whereas the difficulty parameter were drawn from the N(0,1) distribution. Each condition was replicated 200 times. We used an EG design and examined the following four different conditions; (1) The baseline case. (2) A more difficult Y test form (achieved by adding 0.5 to the difficulty parameters in test form Y), (3) More discriminating Y test form (achieved by adding 0.5 to the discrimination parameters in test form Y). (4) More able test taker took test form Y (achieved by assuming the

ability was drawn from N(1,1) for population Q). The examined conditions were compared with respect to the obtained equating transformation and the root mean squared error (RMSE).

The R package *ltm* (Rizopoulos, 2006) was used to generate the dichotomous score responses. The R package *fda* (Ramsay et al., 2022) was used to calculate the landmark registration, the R package *TestGardener* (Ramsay & Li, 2021) was used to calculate the binary optimal scores and the R package *equate* (Albano, 2016) was used to calculate the equipercentile equating.

6.1 Simulation Study Results

We started by examining the RMSE in the four different conditions, and the result is given in Fig. 2. The overall result was that landmark registration yielded lower RMSE within each test score as compared to equipercentile equating. The sum scores had low RMSE also for equipercentile equating except in the higher percentiles. The landmark registration had more similar RMSE over the percentiles than equipercentile equating, even though we only used five landmarks. Only small differences appeared in the four conditions for landmark registration. The equipercenitle equating for sum scores all showed the same pattern, with a sudden rise of RMSE for sum scores at the highest percentiles. The reason is probably due to the low amount of test takers achieving the highest sum scores. For a more discriminating test Y, the RMSE was lower in all four conditions. If a more able test group took test Y the RMSE for the equipercentile equating was much higher for both test scores.

Figure 2 strengthen the pattern observed in the empirical study, i.e. that optimal scores and sum scores differ quite much as they are built differently, i.e. the optimal score depend on which test item is answered correctly and not as in sum scores based on binary items where all items have the same weight. Thus in the following figures we primarily focused on the results of the equating methods within each test score instead of comparing the results between the test scores.

The left part of Fig. 3 gives the equated values for sum scores using equipercentile equating in the four conditions. It is clear that all different conditions affect the equated values quite much. The right part of Fig. 3 displays the equated values for optimal scores using equipercentile equating in the four conditions. In contrast to sum scores, the item discriminations had the largest affect on the equated scores when optimal scores were used but the other conditions did not have any large effect.

Figure 4 displays the equated values at different percentiles using landmark registration for sum scores to the left and optimal scores to the right in the four conditions. Again, item discrimination heavily affect the equated values when optimal scores are used with landmark registration, but all conditions were affected when sum scores were used with landmark registration. Note, the reason percentiles were used for the original scale here as landmark registration are built on percentiles and we wanted to emphasize this.

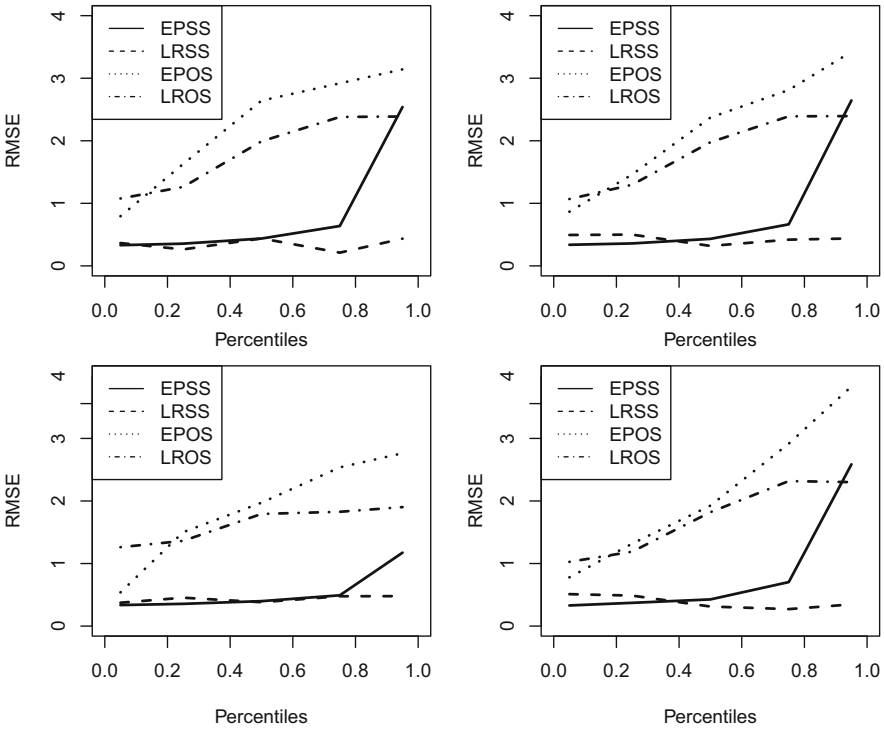


Fig. 2 RMSE for sum scores (SS) and optimal scores (OS) with equipercentile equating (EP) and landmark registration (LR). From top left to right bottom; baseline case, more difficult Y test, more discriminating Y test and more able test takers taking test Y

7 Discussion

In this paper, we have illustrated how to equate test forms scored with sum scores and optimal scores. We also proposed to equate test scores with landmark registration, i.e. to register one CDF to another to reduce variation by using theory from functional data analyses. We compared the proposed equating method with equipercentile equating using sum scores and optimal scores with both real data and a simulation study.

The empirical example support the use of either equipercentile equating or the proposed equating method when we have an EG design. This is good news as optimal scores have been shown to be advantageous for test takers, especially for high achievers (Ramsay & Wiberg, 2017a; Wiberg et al., 2018). To be able to equate test scores from different test versions is important if optimal scores are to be used in practice. One advantage of the proposed method is that the minimum and maximum test scores are kept intact in the equated scores which is not necessary the case for linear equating. It is also flexible as one can decide which percentiles matters

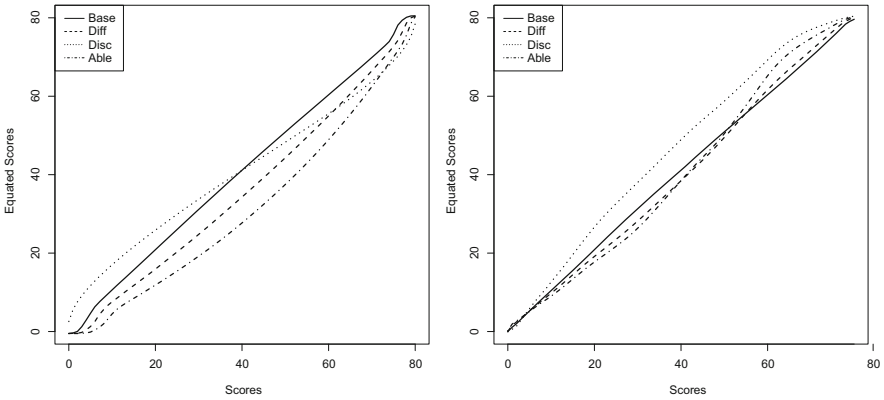


Fig. 3 Equated values for sum scores to the left and optimal scores to the right with equipercentile equating in the four conditions

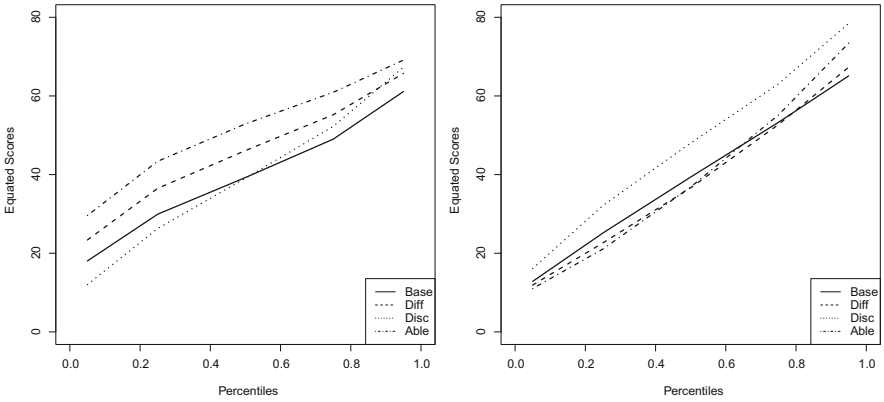


Fig. 4 Equated values for sum scores to the left and optimal scores to the right with landmark registration at different percentiles in the four conditions

more, by setting them as landmarks. It is of course also possible to examine specific percentiles in equipercentile equating but the idea with landmark registration is to put emphasis on certain percentiles. This has potential to be useful when equating test forms where one are interested in putting percentile markers for different grades.

The simulation study showed that the largest differences were between different test scores and only smaller differences were found between the equating methods within the test scores. The RMSE was lowest for landmark registration for sum scores in all examined conditions. The RMSE for sum scores using equipercentile was most affected by change in item discrimination. On contrary, RMSE for optimal scores using equipercentile was most affected by the change in the examinees' ability. An advantage of landmark registration in comparison to equiprecentile equating is the possibility to get stable equated scores even if very few test takers

have scores in the lower or upper score range. The result that landmark registration can be used for sum scores makes it interesting, especially when we have observed test scores which are not evenly distributed on the score scale.

Several research topics can be of interest in the future. First, the proposed method should be examined for the case when we want to equate several test forms. Recently, parametric IRT equating methods have been developed for this purpose (Battauz, 2017) but a drawback is that parametric IRT can be computational challenging. An advantage with optimal scores is they are in general less computational demanding than parametric IRT. As landmark registration is useful in functional data analysis when we have many curves it should be relatively straight forward to examine this further. Second, in our simulation study we generated the scores with IRT, in the future one should consider a larger comparison where one simulate test scores without a parametric model as in e.g. Leoncio et al. (2022). Third, one could also examine test forms which have polytomous items and then one could use optimal scores as described in Ramsay et al. (2020a). Fourth, it would be interesting to examine optimal scores if differential item functioning is present in some or several of the items. Fourth, to make landmark registration more useful one should study how it performs if a nonequivalent groups with anchor test design is used instead of an EG design. This application should be straight forward, as we can use a chained equating approach, possible referred to as chained warping equating, in which the EG procedure is extended as follows. Instead of using a single warping function we need to use a chain of warping functions. First, we register the CDF of test form X to the anchor test CDF in population P, then we register the CDF of the anchor test in population P to the CDF of the anchor test in population Q. Finally, the CDF of the anchor test in population Q is registered to the CDF of test Y. The obtained chained warping function will be similar to chained equipercentile equating but the CDF's are obtained differently. How it would work in comparison to other methods within the NEAT design is however left for future research.

In summary, the proposed equating method and the examination of using equipercentile equating with optimal scores extends the possibilities to use optimal scoring in different test situations. There are however still important topics that needs to be studied in the future in order to make optimal scores a natural choice in standardized testing.

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

References

Albano, A. D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software, 74*(8), 1–36.

Battauz, M. (2017). Multiple equating of separate irt calibrations. *Psychometrika, 82*(3), 610–636.

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

González, J., & Wiberg, M. (2017). *Applying test equating methods using R*. Springer.
Kolen, M., & Brennan, R. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer.
Leoncio, W., Wiberg, M., & Battauz, M. (2022). Evaluating equating transformations in IRT observed-score and kernel equating methods. *Applied Psychological Measurement*. https://doi.org/10.1177/01466216221124087.
Lord, F. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.
Ramsay, J. O. (1996). A geometrical approach to item response theory. *Behaviormetrika, 23*, 3–16.
Ramsay, J. O., & Li, J. (2021). Testgardener: Optimal analysis of test and rating scale data. *Computer software freely available at CRAN*.
Ramsay, J., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). Springer-Verlag.
Ramsay, J. O., & Wiberg, M. (2017a). Breaking through the sum scoring barrier. In *Quantitative psychology – 81st annual meeting of the psychometric society* (pp. 151–158). Springer.
Ramsay, J. O., & Wiberg, M. (2017b). A strategy for replacing sum scoring. *Journal of Educational and Behavioral Statistics, 42*(3), 282–307.
Ramsay, J., Hooker, G., & Graves, S. (2009). *Functional data analysis with R and MATLAB*. Springer.
Ramsay, J. O., Li, J., & Wiberg, M. (2020a). Better rating scale scores with information-based psychometrics. *Psych, 2*(4), 347–369.
Ramsay, J. O., Li, J., & Wiberg, M. (2020b). Full information optimal scoring. *Journal of Educational and Behavioral Statistics, 45*(3), 297–315.
Ramsay, J., Graves, S., & Hooker, G. (2022). fda: Functional data analysis. *Computer software freely available at CRAN*.
Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1–25.
von Davier, A. A., Holland, P., & Thayer, D. (2004). *The kernel method of test equating*. Springer.
Wiberg, M., Ramsay, J. O., & Li, J. (2018). Optimal scores as an alternative to sum scores. In *Quantitative psychology – 82nd annual meeting of the psychometric society* (pp. 1–10). Springer.
Wiberg, M., Ramsay, J. O., & Li, J. (2019). Optimal scores: An alternative to parametric item response theory and sum scores. *Psychometrika, 84*(1), 310–322.

Pauci sed boni: An Item Response Theory Approach for Shortening Tests

Ottavia M. Epifania, Pasquale Anselmi, and Egidio Robusto

Abstract Item Response Theory (IRT) is the theoretical framework often used for shortening tests. This contribution presents a new IRT-based item selection procedure which is meant for this purpose. This procedure is based on the information that each item provides in respect to different trait levels of interest (denoted as θ targets), which are obtained by segmenting the latent trait in either equal or unequal intervals. In a simulation study, the performance of the new procedure was compared with that of the typical IRT procedure and of a random selection of the items. The new procedure outperformed the other two in recovering central and peripheral regions of the latent trait continuum, particularly when the short test forms consisted of fewer items. Despite this study highlighted the potentiality of the new item selection procedure for developing short test forms, work is still needed.

Keywords Item response theory · Static short test form · Information · Assessment precision

1 Introduction

Tests can be efficiently shortened by considering the information at the item level provided by Item Response Theory (IRT) models (see, e.g., Colledani, Anselmi, & Robusto, 2021; Chiesi, Lau, & Saklofske, 2020; Choi, Reise, Pilkonis, Hays, & Cella, 2010; Silvia, 2021; Edelen & Reeve, 2007). According to IRT models, the probability of observing a correct response on an item is a function of

The authors have no conflict of interest to disclose.
This research was funded by "BIRD SID assegni 2020" with code C59C20000050005.

O. M. Epifania (✉) · P. Anselmi · E. Robusto
Department of Philosophy, Sociology, Education, and Applied Psychology, University of Padova (IT), Padova, Italy
e-mail: ottavia.epifania@unipd.it; pasquale.anselmi@unipd.it; egidio.robusto@unipd.it

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_7

the characteristics of the person taking the test (i.e., the latent trait) and the characteristics of the item. Additionally, IRT models allow for obtaining information on the precision with which each item assesses different levels of the latent trait. This information is of particular relevance for the development of short test forms (STFs). Generally, IRT models can be used to develop STFs following one of two approaches. One approach is used for obtaining adaptive STFs where the items administered to each person are adaptively selected during the test administration. Computerized adaptive testing (CAT) procedure results in adaptive STFs that can provide precise assessments of the latent trait of the test-takers while including the least number of items (e.g., Drasgow & Olson-Buchanan, 1999; Magis & Barrada, 2017). In CAT procedures, each adaptive STF can be different according to the level of the latent trait of each individual. As such, each individual can be administered with different subsets of items, which are adaptively selected according to the level of the latent trait of the test taker. Although having a STF tailored to the specific level of the latent trait of each individual is optimal in terms of information, it is not ideal in specific contexts such as job recruitment or college admission, where the different subsets of items might raise fairness issues. The other approach results in static STFs, and it is the one usually employed for shortening tests in an IRT framework (see, e.g., Colledani et al., 2019, 2018, 2021; Chiesi et al., 2020; Silvia, 2021). In static STFs, the most informative items are chosen, irrespective of the latent trait of the test-takers. As such, all test-takers are administered with the same subset of items. Since the static STFs are not tailored to any specific level of the latent trait, they might require a larger number of items for a precise assessment of different levels of the latent trait than adaptive STFs obtained with CAT procedures. Other approaches exist for shortening tests in an IRT-based framework, such as those based on the two stage semi-adaptive branching (Belov & Armstrong, 2008), according to which a common item is asked first and, based on the response to that item, a different subset of items is administered. Differently from CAT, where each item is chosen according to the response to the previous item, in semi-adaptive branching procedures the response to the first item determines the subset of items to be administered.

In this contribution, we focus on the development of static STFs. Specifically, we aim to obtain a procedure that strives for combining the main advantages of the adaptive STFs (i.e., being tailored to specific trait levels of interest) with those of the static STFs (i.e., being equal for all respondents). The following section gives an overview of the IRT model used for the application of the proposed procedure, along with an introduction to the item and test information functions. In Sect. 3, the typical procedure for shortening tests in an IRT framework and the procedure introduced in this study are presented. The results of a simulation study are presented in Sect. 4. A discussion of the results of the simulation study close the argumentation.

2 Item Response Theory and Information Functions

The procedure presented in this contribution is based on the 2-parameter logistic model (2-PL Birnbaum, 1968) for dichotomous responses. However, the proposed approach is general enough to be used with other IRT models for both dichotomous and polytomous responses. In the 2-PL model, the probability of a correct response to item i by person p is formalized as:

$$P(x_{pi} = 1|\theta_p, b_i, a_i) = \frac{exp[a_i(\theta_p - b_i)]}{1 + exp[a_i(\theta_p - b_i)]} \tag{1}$$

where θ_p is the level of the latent trait of person p, and b_i and a_i are the difficulty and discrimination parameters of item i, respectively. The difficulty parameter b describes the location of item i on the latent trait. The discrimination parameter a describes the strength with which item i is linked to the latent trait (i.e., the ability of the item to discriminate between respondents with high and low levels of the latent trait).

The item information function (*IIF*) informs about the precision with which an item measures the latent trait. In the 2PL model, the IIF_i of item i is obtained as:

$$IIF_i = a_i^2[P(\theta)(1 - P(\theta))], \tag{2}$$

where $P(\theta)$ is the probability of a person with a certain θ of responding correctly to item i, and $1 - P(\theta)$ is their probability of responding incorrectly. The higher the discrimination parameter of item i, the higher IIF_i. The *IIF* reaches its maximum in proximity of the location of the item on the latent trait (i.e., item i is most informative when the trait level of the respondent matches the difficulty parameter b_i). By summing up the *IIF* of all the items, a measure of the overall precision of the test in measuring the latent trait is obtained (i.e., test information function, $TIF = \sum_{i=1}^{I} IIF_i$).

3 Item Selection Procedures

3.1 Benchmark Procedure

The benchmark procedure (BP) is the typical IRT procedure for shortening tests. The N items with the highest *IIF*s are selected from the items of the full-length test to be included in the static STF, where N is the desired length for the static STF.

$$\begin{bmatrix} 0.02 & 0.03 & 0.03 \\ 0.02 & 0.01 & 0.01 \\ 0.14 & 0.29 & 0.31 \\ 0.03 & 0.03 & 0.04 \\ 0.05 & 0.06 & 0.06 \\ 0.12 & 0.12 & 0.09 \\ 0.01 & \mathbf{0.51} & 0.05 \\ 0.17 & 0.04 & 0.01 \\ 0.05 & 0.05 & 0.04 \\ 0.33 & 0.07 & 0.10 \end{bmatrix} \quad \begin{bmatrix} 0.02 & 0.03 & 0.03 \\ 0.02 & 0.01 & 0.01 \\ 0.14 & 0.29 & 0.31 \\ 0.03 & 0.03 & 0.04 \\ 0.05 & 0.06 & 0.06 \\ 0.12 & 0.12 & 0.09 \\ 0.01 & 0.51 & 0.05 \\ 0.17 & 0.04 & 0.01 \\ 0.05 & 0.05 & 0.04 \\ \mathbf{0.33} & 0.07 & 0.10 \end{bmatrix} \quad \begin{bmatrix} 0.02 & 0.03 & 0.03 \\ 0.02 & 0.01 & 0.01 \\ 0.14 & 0.29 & \mathbf{0.31} \\ 0.03 & 0.03 & 0.04 \\ 0.05 & 0.06 & 0.06 \\ 0.12 & 0.12 & 0.09 \\ 0.01 & 0.51 & 0.05 \\ 0.17 & 0.04 & 0.01 \\ 0.05 & 0.05 & 0.04 \\ 0.33 & 0.07 & 0.10 \end{bmatrix}$$

Fig. 1 Illustration of the θ' procedure for developing a static STF composed of $N = 3$ items from a full length test composed of $J = 10$ items. Only the gray rows/columns are considered for the item selection at each iteration. Left, central, and right panels illustrate the **IIF** matrix at iteration 1, 2, 3, respectively

3.2 Procedure Based on θ Targets

The procedure is based on the definition of trait levels of interest (i.e., denoted as θ targets, θ's), which are the levels of the latent trait on which the static STF focuses the most. The items that best assess the trait levels of interest (i.e., optimal items) are included in the static STF. In what follows, this procedure is referred to as θ-target procedure.

The latent trait is segmented into N θ's, where N is the number of items to be included in the static STF. The *IIF*s of each of the J items composing the full-length test are computed for each θ' and they are arranged in a $J \times N$ matrix **IIF**. The procedure iterates from 0 to $N - 1$. At each iteration, the item with the highest *IIF* in **IIF** is selected for the inclusion in the static STF. Once an item has been selected for a specific θ', the row corresponding to that item and the column corresponding to that θ' are not available anymore for item selection at the subsequent iteration. As soon as an optimal item (i.e., the item with the highest *IIF*) has been identified for each θ', the procedure stops.

The following example illustrates the procedure based on θ' for creating a static STF composed of $N = 3$ items from a full length test composed of $J = 10$ items (Fig. 1).

At the first iteration, all the rows and the columns of **IIF** are considered for the item selection (gray area in the left matrix). The cell with the highest *IIF* is **IIF**$(6, 2)$. Row 6 of **IIF** corresponds to item 6 while column 2 of **IIF** corresponds to θ'_2. Item 6 is the best item for evaluating θ'_2 and it is selected for the inclusion in the STF.

Since item 6 has already been selected as the optimal item for θ'_2, the sixth row and the second column of **IIF** are not available for the item selection at the second iteration (central matrix). The highest *IIF* is in **IIF**$(10, 1)$. Item 10 is the optimal item for θ'_1, and it is selected for the inclusion in the STF.

At the third iteration, the rows corresponding to items 6 and 10 and the columns corresponding to θ'_2 and θ'_1 are no longer available for the item selection (right matrix). The highest *IIF* is in **IIF**$(3, 3)$. Being the optimal item for θ'_3, item 3 is

selected for the inclusion in the STF. Since the number of selected items is equal to N, the procedure ends.

In this contribution, two methods for selecting the θ's are presented, according to which the latent trait can be either clustered into N clusters or it can segmented into $N + 1$ intervals of equal width. In the former case, the latent trait is segmented into unequal intervals and the centroids c_n of the N clusters are the θ's (unequal intervals procedure, UIP). In the latter case, the latent trait is segmented into equal intervals, and the N central values of the $N + 1$ intervals are the θ's (equal intervals procedure, EIP).

3.3 Comparison Between θ-Target Procedure and Benchmark Procedure

Both the BP and the θ-target procedure aim at obtaining static STFs. While the BP procedure selects the item according to their information functions, irrespective of the position of the items on the latent trait, the θ-target procedure selects the most informative item for each of the identified θ targets. Unless the item selection in the BP procedure is supported by a visual inspection of the latent trait (e.g., Chiesi et al., 2020), the items are selected only according to their information functions, regardless of their location on the latent trait. As such, the items with locations that match the most common levels of the latent trait of the respondents have a higher probability of being selected than the items with locations matching the least common levels of the latent trait. The risk associated with such a procedure is that the static STF might not precisely assess the respondents with extreme levels of latent trait, this resulting in biased estimates of the latent trait (see, e.g., Feuerstahler, 2018)

In the θ-target procedure, the item selection considers the information that each item provides in respect to the trait levels of interest (i.e., θ targets). Specifically, the most informative item for each θ target is selected in the iterative procedure. Since the θ targets are spread along the entire latent trait and the most informative item has been selected for each of them, the resulting static STFs should be able to provide a precise and reliable assessment of both the most dense and the least dense regions of the latent trait. In this sense, the item selection is tailored to each θ target. Therefore, this procedure is expected to maximize the information and the assessment precision across all respondents, including those with extreme levels of the latent trait.

4 Method

The performance of the BP and that of the procedures based on θ's (i.e., UIP and EIP) are compared for the development of STFs composed of 10, 30, 50, 70, and 90

items obtained from a full-length test composed of 100 items. A random selection of items (random procedure, RP) from the full length test is considered as well. The performance of the procedures is evaluated in terms of *TIF* and overall coverage of the latent trait.

The latent traits of 1000 respondents were simulated from a normal distribution $\mathcal{N}(0, 1)$. Item difficulty parameters b were simulated from a uniform distribution $\mathcal{U}(-3, 3)$. Since it is not uncommon to find most discriminative items concentrated in the central region of the latent trait (i.e., medium difficulty levels, Azzopardi & Azzopardi, 2019; Sim & Rasiah, 2006), medium to highly discriminative items ($a > .64$ Baker & Kim, 2017) were assigned to the items with medium levels of difficulty, while lowly discriminative items ($a \leq .64$ Baker & Kim, 2017) were assigned to items with most extreme levels of difficulty. The item discrimination parameters were simulated from a χ^2 distribution with 2 degrees of freedom, and the values were multiplied by a constant (4.3) to obtain plausible item discrimination parameters. Difficulty and discrimination parameters were simulated for 100 items.

5 Results

The overall information of the STFs of different length is reported in Fig. 2. Benchmark procedure (BP) resulted in the highest information irrespective of the length of the STFs. As the number of items increased (i.e., 50-item STF), the performance of BP and UIP tended to be similar. EIP resulted in lower overall information than both BP and UIP. RP resulted in the lowest information.

The TIFs of all static STFs are reported in Fig. 3. Irrespective of the number of items included in the STF, the BP (solid line) always resulted in the highest TIFs, while EIP (dot-dashed line) resulted in the lowest TIFs (after RP, dotted line). However, the peripheral regions of the latent trait were more precisely assessed by EIP and UIP than by BP. This is particularly noticeable for shorter static STFs.

6 Discussion

In this contribution, we presented a new procedure for the development of static STFs in an IRT framework. By tailoring the item selection to specific and fixed levels of the latent trait, the procedure is supposed to be a trade-off between the procedures used for adaptive testing and those typically used for the development of static STFs.

The results highlighted the potential of the new procedure for developing STFs able to efficiently and precisely measure also the peripheral (i.e., least common) regions of the latent trait. The assessment precision of the extreme regions of the latent trait might be of particular use when non-normally distributed latent traits are

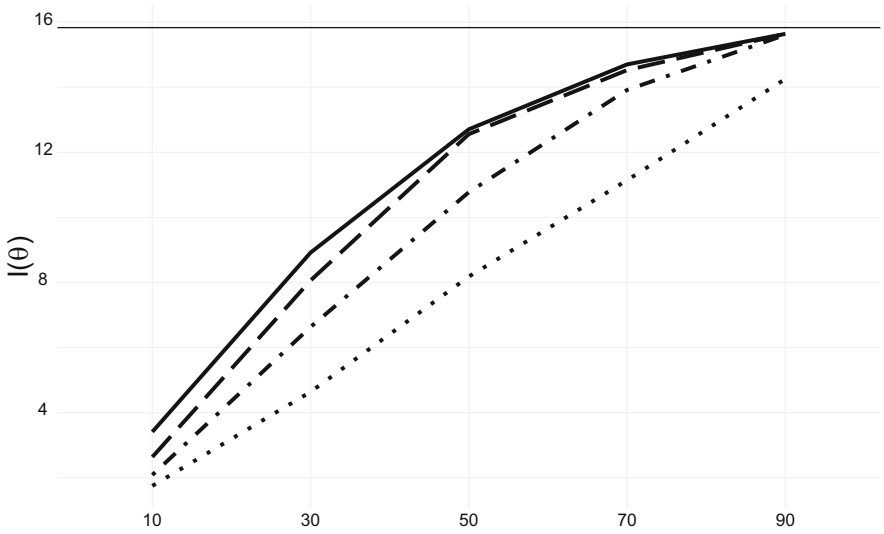


Fig. 2 Overall information of the short test forms. Solid line: Benchmark procedure. Long-dashed line: Unequal intervals procedure. Dot-dashed line: Equal intervals procedure. Dotted line: Random procedure

considered, such as for the assessment of medical health outcomes (see, e.g., Smits et al., 2020) or for the assessment of clinically relevant constructs among the general population (e.g., Anselmi et al., 2022; Colledani et al., 2021), where the latent trait might present strong floor or ceilings effects. If a STF is to be developed for the assessment of such constructs, it should be able to provide a reliable assessment of the most common regions of the latent trait where the majority of the respondents are located and also of the extreme regions of the latent trait. This would allow for precisely assessing the majority of the respondents falling in the most dense regions of the latent trait and for identifying the respondents with extreme levels of the latent trait by administering the same STF. Moreover, the θ-target procedure might be useful for developing STFs focused on specific trait levels, such as for the diagnostic screening of clinically relevant topics in the general population. In such instances, the trait levels of interest (hence the θ targets) would be those around the cut-off level used for identifying potentially problematic respondents. This would ensure to obtain a STF able to adequately discriminate between respondents with trait levels over and below the cut-off level.

Since items with medium difficulty levels tend to be also the most discriminative ones (e.g., Azzopardi & Azzopardi, 2019; Sim & Rasiah, 2006), the item parameters in this study were simulated such that the most discriminative items could be concentrated in one region of the latent trait. However, the performance of the proposed procedure should be tested and compared with other procedures also considering situations where highly and lowly discriminative items are equally spread throughout the entire latent trait.

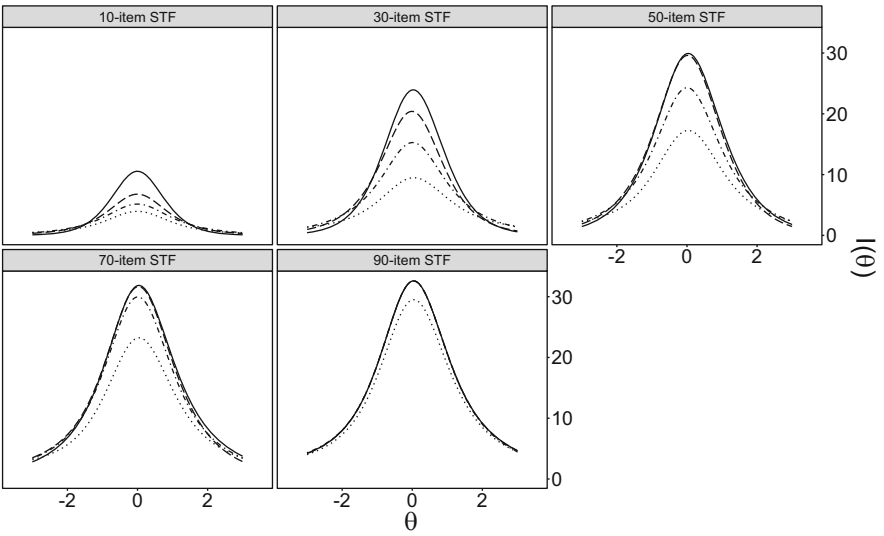


Fig. 3 Test information functions of all short test forms. Solid line: Benchmark procedure. Long-dashed line: Unequal intervals procedure. Dot-dashed line: Equal intervals procedure. Dotted line: Random procedure

The lack of a comparison between the assessment precision of STFs obtained with the CAT procedure and that of STFs obtained with the θ-target procedure is a limitation of the study. Future studies should compare the assessment precision of the STFs obtained with all the IRT-based procedures used for shortening tests, including the one presented in this contribution. In the θ-target procedure, the item selection is tailored to the identified θ targets, similarly to the underlying logic of CAT procedures where the selected items are tailored to the specific θ level of each respondent. Differently from CAT, in the θ target procedure the items are selected to maximize the information across all respondents by administering the same subset of items, similarly to what it is done in the typical IRT procedure for shortening tests. As such, the θ-target procedure can be considered as a sort of middle ground between the typical IRT procedure and the CAT procedure. Following this idea, the assessment precision of the STFs obtained with the θ-target procedure should be better than that of the STFs obtained with the typical IRT procedure but worse than that of the STFs obtained with the CAT procedure. Nonetheless, the adaptive STFs obtained with the CAT procedure hinder the comparability between respondents, given that they are administered with different subsets of item. This potential issue is overcome in the θ-target procedure.

In conclusion, the presented approach showed its feasibility for the development of STFs in an IRT framework. However, work is still needed to further understand the applicability and advantages of the new procedure considering different scenarios.

References

Anselmi, P., Colledani, D., Andreotti, A., Robusto, E., Fabbris, L., Vian, P., et al. (2022). An item response theory-based scoring of the south oaks gambling screen-revised adolescents. *Assessment, 29*(7), 1381–1391. https://doi.org/10.1177/10731911211017657.

Azzopardi, M., & Azzopardi, C. (2019). Relationship between item difficulty level and item discrimination in biology final examinations. *Education and New Developments*. https://doi.org/10.36315/2019v2end001.

Baker, F. B., & Kim, S.-H. (2017). *The basics of item response theory using R*. Springer.

Belov, D. I., & Armstrong, R. D. (2008). A Monte Carlo approach to the design, assembly, and evaluation of multistage adaptive tests. *Applied Psychological Measurement, 32*(2), 119–137. https://doi.org/10.1177/0146621606297308

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

Chiesi, F., Lau, C., & Saklofske, D. H. (2020). A revised short version of the compassionate love scale for humanity (CLS-H-SF): Evidence from item response theory analyses and validity testing. *BMC Psychology, 8*(1), 1–9. https://doi.org/10.1186/s40359-020-0386-9.

Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research, 19*(1), 125–136.

Colledani, D., Robusto, E., & Anselmi, P. (2018). Development of a new abbreviated form of the junior Eysenck personality questionnaire-revised. *Personality and Individual Differences*, *120*, 159–165. https://doi.org/10.1016/j.paid.2017.08.037

Colledani, D., Anselmi, P., & Robusto, E. (2019). Using multidimensional item response theory to develop an abbreviated form of the italian version of eysenck's ive questionnaire. *Personality and Individual Differences*, *142*, 45–52.

Colledani, D., Anselmi, P., & Robusto, E. (2021). Cross-cultural validation of a new abbreviated version of the EPQ-R. *TPM: Testing, Psychometrics, Methodology in Applied Psychology*, *28* (2021). https://doi.org/10.4473/TPM28.3.6

Drasgow, F., & Olson-Buchanan, J. B. (1999). *Innovations in computerized assessment*. Psychology Press.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*(1), 5–18. https://doi.org/10.1007/s11136-007-9198-0

Feuerstahler, L. M. (2018). Sources of error in IRT trait estimation. *Applied Psychological Measurement*, *42*(5), 359–375. https://doi.org/10.1177/0146621617733955

Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. *Journal of Statistical Software*, *76*, 1–19. https://doi.org/10.18637/jss.v076.c01

Silvia, P. J. (2021). The self-reflection and insight scale: Applying item response theory to craft an efficient short form. *Current Psychology*, 1–11. https://doi.org/10.1007/s12144-020-01299-7

Sim, S.-M., & Rasiah, R. I. (2006). Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. *Annals-Academy of Medicine Singapore*, *35*(2), 67.

Smits, N., Öğreden, O., Garnier-Villarreal, M., Terwee, C. B., & Chalmers, R. P. (2020). A study of alternative approaches to non-normal latent trait distributions in item response theory models used for health outcome measurement. *Statistical Methods in Medical Research*, *29*(4), 1030–1048. https://doi.org/10.1177/09622802209076

Limited Utility of Small-Variance Priors to Detect Local Misspecification in Bayesian Structural Equation Models

Terrence D. Jorgensen and Mauricio Garnier-Villarreal

Abstract In a highly influential paper on current practice in Bayesian structural equation modeling (BSEM), Muthén and Asparouhov (Psychol Methods 17:313–335, 2012) proposed using small-variance priors to constrain non-target parameters to be close to (rather than exactly) zero, with the "side product" (p. 313) that the posterior distributions of such nontarget parameters could be used analogous to modification indices. This chapter presents 2 simulation studies of their utility, in the context of (a) constraining cross-loadings to be nearly zero and (b) constraining factor loadings and intercepts to be equivalent across groups or occasions. The first study reinforced earlier findings that small-variance priors can prevent detecting important misspecifications (i.e., global-fit indices indicate better fit as priors become less restrictive). In contrast, these local indicators have greater power to detect invalid constraints when priors are less restrictive. Study 2 revealed similar patterns in the context of detecting invalid equality constraints and showed limited utility of small-variance priors over modification indices under maximum-likelihood estimation. Our advice is to evaluate global fit in BSEM without small-variance priors, and only when hypothesized models are rejected, utilize small-variance priors to search for clues about possible respecification. We recommend exploring other tools for local-fit evaluation in BSEM, which might detect misspecifications without introducing additional complications of small-variance priors (e.g., propagation of bias).

Keywords Bayesian · Structural equation modeling · Measurement invariance · Differential item functioning · Modification indices

T. D. Jorgensen (✉)
Universiteit van Amsterdam, Amsterdam, the Netherlands
e-mail: t.d.jorgensen@uva.nl

M. Garnier-Villarreal
Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_8

1 Introduction

Bayesian structural equation modeling (BSEM) has recently received substantial attention within psychology and the social sciences as an increasingly viable alternative to traditional frequentist SEM techniques, such as maximum likelihood (ML) estimation. Several tools are available to evaluate global (mis)fit of a BSEM, such as posterior predictive model checking (PPMC; Gelman et al., 1996), for which a posterior predictive *p* value (PPP) can be calculated that is analogous to the *p* value of a SEM's χ^2 statistic, which tests the null hypothesis (H_0) that a SEM perfectly represents the true data-generating process. Approximate global fit of a BSEM can be evaluated using SRMR (Levy, 2011) or χ^2-based fit indices analogous to those under maximum likelihood estimation (MLE), on the condition that the BSEM uses uninformative priors during Markov chain Monte Carlo (MCMC) estimation (Garnier-Villarreal & Jorgensen, 2020). Although PPP or fit indices may indicate poor model fit, they cannot provide clues about the specific source(s) of misspecification.

In a highly influential paper, Muthén and Asparouhov (2012) proposed using small-variance priors to constrain non-target parameters to be close to zero, as a less-restrictive alternative to fixing such parameters to exactly zero. The Bayesian credible intervals (BCI; interval estimates analogous to confidence intervals of frequentist estimators) for nontarget parameter estimates (constrained to be small) can be used to indicate local sources of misspecification. They suggested that "[the sensitivity of nontarget parameters] be used in line with modification indices [in MLE] to free parameters for which the credibility interval does not cover zero" (Muthén & Asparouhov, 2012, pp. 316–317), noting the advantage over modification indices in that BCIs for all parameters can be obtained simultaneously, preventing the problem of sequentially modifying one parameter at a time under ML estimation. The goal of this paper is to evaluate their proposal in the context of (a) cross-loadings in single-group SEM and (b) equality constraints on loadings (i.e., measurement equivalence) using Monte Carlo simulations.

2 Study 1: Priors for Approximately-Zero Constraints

This study was part of an investigation of PPP's frequency properties, so the Method details correspond to those published by Jorgensen et al. (2019). We focus only on normal-data conditions here because patterns of results for ordinal data were largely similar, although power decreased with fewer categories.

2.1 Method

Using the MONTECARLO command in M*plus* (version 6.11 for Linux; Muthén & Muthén, 2012), we simulated a two-factor CFA with three indicators per factor. In each of the four population models, factors were standard normal ($\mu = 0$, $\sigma = 1$), with a factor correlation $\psi_{21} = 0.25$, factor loadings $\lambda = 0.7$, indicator intercepts $= 0$, and indicator residual variances $\theta = 0.51$; thus, indicators had unit variance. To vary levels of misspecification of the analysis model, the third indicator of the first factor was specified to have a cross-loading on the second factor (λ_{32}) in the population. The magnitude of λ_{32} was 0.0, 0.2, 0.5, or 0.7 in the population, but was constrained to be close to zero in the analysis model using informative priors (see next paragraph). For ease of interpretation, we refer to $\lambda_{32} = 0.2$ as minor misspecification (using $\alpha = .05$, the ML χ^2 test has 80% power when $N > 500$, RMSEA = 0.06, SRMR = 0.03, CFI = 0.98), $\lambda_{32} = 0.5$ as severe misspecification (80% power when $N > 150$, RMSEA = 0.12, SRMR = 0.07, CFI = 0.92), and $\lambda_{32} = 0.7$ as very severe misspecification (80% power when $N > 100$, RMSEA = 0.14, SRMR = 0.07, CFI = 0.89).

In the analysis model, we specified noninformative priors for all target parameters (primary loadings, residual variances, and the factor covariance) using M*plus* defaults—for example, factor loadings ~ $N(\mu = 0, \sigma^2 =$ "infinity"). For all cross-loadings, we specified normally distributed priors with four levels of informative variance, chosen to correspond approximately with the prior belief in a 95% probability that the cross-loadings are within approximately ±0.01, ±0.10, ±0.20, or ± 0.30 of zero (i.e., $\sigma = 0.005$, 0.05, 0.10, and 0.15, or equivalently $\sigma^2 = 0.000025$, 0.0025, 0.01, and 0.0225). In each condition, sample sizes of $N = 50$–500 were drawn in increments of 25, along with an asymptotic condition of $N = 1000$. We generated 200 samples from each of 320 conditions (20 sample sizes, four levels of CL, and four prior variances) with normally distributed indicators.

We kept 100,000 iterations from the MCMC chains after thinning every 100th iteration. Over 99% of models converged on a proper solution, yielding 63,480 (out of 64,000) PPP values for analysis. Convergence was evaluated using Gelman and Rubin's (1992) potential scale reduction factor ("R-hat" < 1.1). Convergence in each condition was at least 98% except when sample size was small ($N < 100$) and CL was large ($\lambda_{32} > 0.5$). The smallest convergence rate was 82% ($N = 50$, $\lambda_{32} = 0.7$). Nonconverged solutions were omitted from Results. Nontarget cross-loadings were considered significantly different from 0 when their 95% BCI excluded 0.

2.2 Results

Whereas the power to reject an inappropriate model increased as prior variance decreased (negative association) when using PPP as an indicator of global misfit (see Jorgensen et al., 2019, for details), the power to detect local sources of misfit

(here, the neglected parameter λ_{32}) increased as prior variance increased (a positive association). Figure 1 depicts how often λ_{32} was detected as significantly different from 0. As would be expected, λ_{32} was never estimated to be significantly greater than 0 when it was in fact 0 in the population, and was very seldom estimated to be significant when it was only 0.2 in the population. As might also be expected, using the most restrictive priors—which yielded the greatest power of PPP to detect misspecification—λ_{32} was never estimated to be significantly greater than 0. Power was only adequate when the neglected parameter was severe ($\lambda_{32} = 0.5$ or 0.7). When the prior variance was reasonably informative (95% CI within ± 0.10 of 0), adequate power ($\geq 80\%$) to detect the neglected cross-loading (λ_{32}) was found for $N > 400$, and for $N > 300$ when priors were less informative (95% CI within ± 0.20 or $\pm$ 0.30 of 0).

We were also interested in the degree to which the neglected cross-loading would affect other parameters estimates in the model. Related cross-loadings (first and second indicators of the first factor, which did not cross-load onto the second factor in the population) were sometimes detected to be significantly different from 0 (although less frequently than the actual neglected cross-loading), and the factor correlation grew increasingly biased. Investigating the average parameter estimates for normal-data conditions in Table 1 (collapsed across sample size, which had no

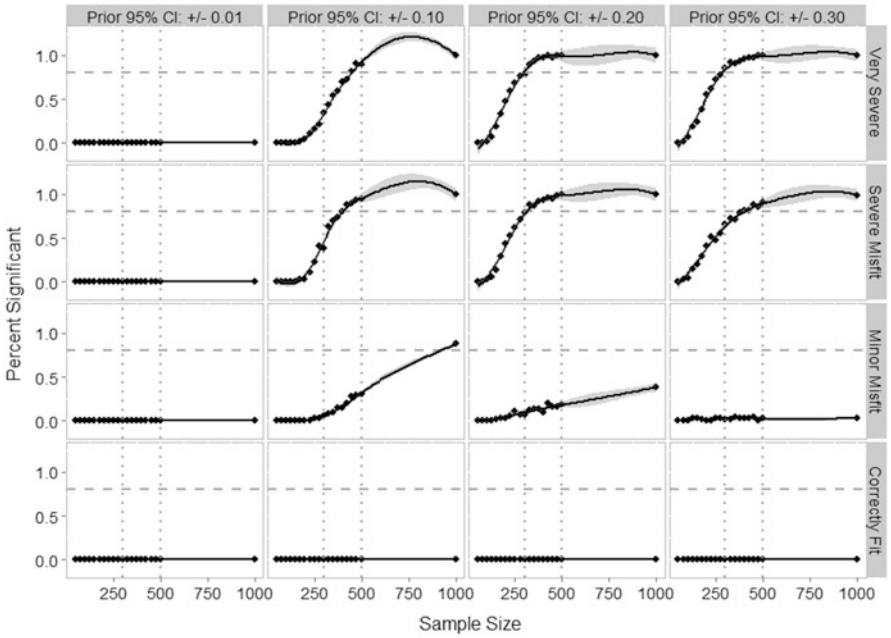


Fig. 1 Rejection rates for neglected cross-loading (λ_{32}) as a function of sample size, plotted separately across conditions of varying priors and magnitude of neglected cross-loading (λ_{32}). Dashed horizontal line provided for reference at 80% power, and dotted vertical lines at $N = 300$ and 500 provided for reference when judging sample sizes necessary for adequate power

Table 1 Effect of neglected cross-loading (λ_{32}) on estimates of related cross-loadings and factor correlation

Prior 95% CI	Population λ_{32}	$\hat{\lambda}_{32}$	$\hat{\lambda}_{12}$	$\hat{\lambda}_{22}$	$\hat{\psi}_{21}$
±0.01	0.0	0.000	0.000	0.000	0.247
	0.2	0.001	−0.001	−0.001	0.332
	0.5	0.001	−0.001	−0.001	0.448
	0.7	0.001	−0.001	−0.001	0.518
±0.10	0.0	0.000	0.001	0.001	0.244
	0.2	0.050	−0.027	−0.026	0.330
	0.5	0.088	−0.058	−0.057	0.438
	0.7	0.087	−0.068	−0.068	0.499
±0.20	0.0	0.002	0.002	0.001	0.243
	0.2	0.091	−0.045	−0.046	0.327
	0.5	0.184	−0.111	−0.110	0.433
	0.7	0.211	−0.145	−0.145	0.492
±0.30	0.0	0.002	0.002	0.001	0.242
	0.2	0.109	−0.055	−0.053	0.329
	0.5	0.234	−0.135	−0.135	0.433
	0.7	0.285	−0.187	−0.186	0.491

effect on the point estimates) reveals that as λ_{32} increased, (a) the average estimates of related cross-loadings decreased, although with less magnitude than the neglected λ_{32}, and (b) the average estimate of the factor correlation became greater than its true value (0.25). Note that although there would seldom be any indication (i.e., low power) that the pattern is significant when the neglected cross-loading is only minor ($\lambda_{32} = 0.2$), such a small neglected parameter estimate still results in a unacceptably biased factor correlation (relative bias = [0.33–0.25] / 0.25 = 0.32), according to Hoogland and Boomsma's (1998) criterion (< 0.05).

To verify that such bias would also occur using MLE, we simulated a single large sample (N = 10,000) from the population with $\lambda_{32} = 0.7$, and fit a model to that data in which all cross-loadings were fixed to 0. This yielded the same negative bias in the related cross-loadings and the same positive bias in the factor correlation. Modification indices indicated that fit would be significantly improved by freeing not only the true omitted cross-loading, but also by freeing other cross-loadings and residual correlations. Freeing only the true omitted cross-loading eliminated bias in any estimates.

3 Study 2: Priors for Approximate Equality Constraints

This is a subset of unpublished results from a dissertation project (Jorgensen, 2015). When evaluating measurement equivalence across contexts (e.g., different populations or occasions), small-variance priors can be specified for parameters that

represent differential item/indicator functioning (DIF), allowing for approximate rather than exact invariance. Although priors can now be easily specified for functions of parameters in M*plus* (Muthén & Muthén, 2012) and `blavaan`[1] (Merkle & Rosseel, 2018), this study was manually programmed in 2014 using Stan (Carpenter et al., 2017).

3.1 Method

Figure 2 represents the data-generating 1-factor SEMs for Study 2. In addition to type of invariance (groups vs. occasions), we manipulated total $N = 200$, 300, 400, 600, or 800 (balanced group sizes) and priors for DIF parameters ~ $N(\mu = 0$, $\sigma = 0.05$ or 0.10; i.e., 95% probability that $\Delta\lambda$ or $\Delta\tau$ fell within ± 0.10 or within ± 0.20, respectively). For longitudinal SEM, the autocorrelation for common and unique factors are indicated by the dashed line representing the factor correlation.

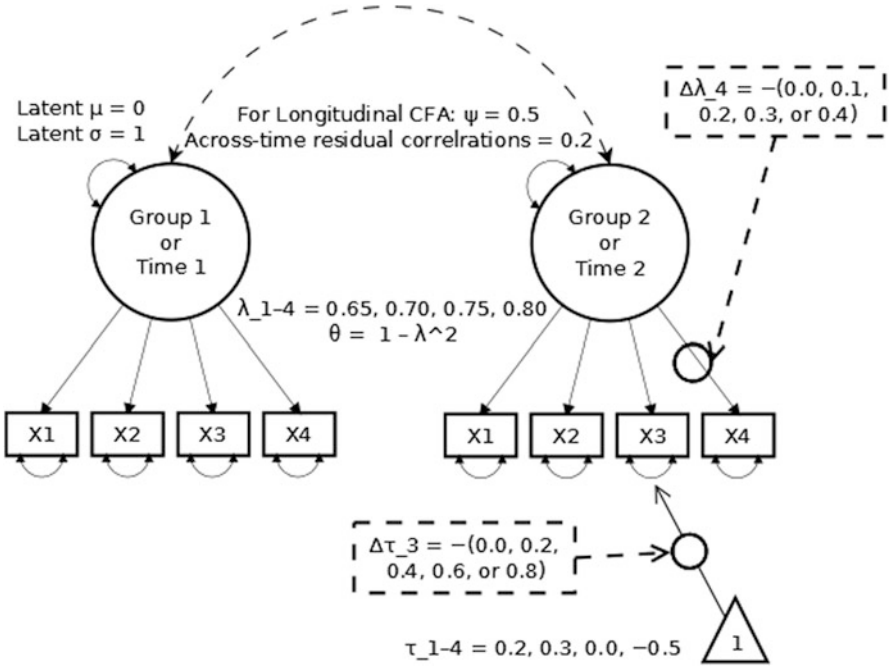


Fig. 2 Population model(s) for data generation in Study 2. Solid lines represent population characteristics that are constant across all conditions, whereas dashed lines represent varying conditions described in the dashed textboxes

[1] See https://ecmerkle.github.io/blavaan/articles/invariance.html for example syntax.

One of 5 effect sizes for DIF (see dashed boxes) were simultaneously added to Item 4's loadings and Item 3's intercept, yielding 5 DIF conditions. We generated 500 samples from each population.

Whereas M*plus* (Muthén & Muthén, 2012) uses Gibbs sampling, Stan (Carpenter et al., 2017) uses a modified Hamiltonian Monte Carlo algorithm called the no U-turn sampler (NUTS), which has efficiency advantages over Gibbs sampling. After 1000 burn-in iterations on each of three chains, we saved 1000 post-burn-in samples per chain. We fit models representing approximate metric invariance (Model 1), approximate full scalar invariance (Model 2b), and approximate partial scalar invariance (Model 2f). Model 2b represents a "backward" specification search, in which DIF is tested by releasing constraints from a fully restricted model. Model 2f represents a "forward" specification search, which proceeds from the least constrained configural model and applies more restrictive constraints. This is not strictly necessary in BSEM because all DIF parameters can be evaluated simultaneously, but it allows comparison of Muthén and Asparouhov's (2012) proposed approach to the traditional use of modification indices in ML estimation (using `lavaan`; Rosseel, 2012).

3.2 Results

Convergence was nearly 100% for Models 2b and 2f, but nonconvergence of Model 1 increased with N, particularly with less informative priors. When the prior $\sigma = 0.05$, convergence dropped from 100% when $N = 200$ to 50% when $N = 800$. When the prior $\sigma = 0.10$, convergence dropped from 100% when $N = 200$ to 25% when $N = 800$. In all conditions, there were > 100 converged results, and collapsing across conditions with little impact (e.g., no substantial differences between multigroup and longitudinal models) increased the Monte Carlo sample sizes used to draw conclusions.

Similar to Study 1, using small-variance priors on substantially nonzero parameters induced bias in other DIF parameters (which were truly zero in the population). Estimated DIF for DIF-free items appeared to counterbalance the invalidly constrained (truly nonzero) DIF parameter, and the effect was stronger in larger samples (see Fig. 3 for estimated DIF in intercepts). Furthermore, estimated parameters (posterior means) of latent variables were systematically biased by using small-variance priors on substantially nonzero parameters. In this case, latent means were biased more negatively as $\Delta\tau_4$ increased, more so in Model 2f (which correctly allowed for DIF in λ_4) than Model 2b (which invalidly constrained λ_4). Surprisingly, less restrictive priors exacerbated the situation: allowing the true DIF to be more negative did not alleviate the truly DIF-free estimates, which were also more positive. Similar results were found for DIF in factor loadings and how that biases estimated latent variance in the second group/occasion (see Jorgensen, 2015, Part III). Patterns were similar but more extreme using ML estimation, when Models

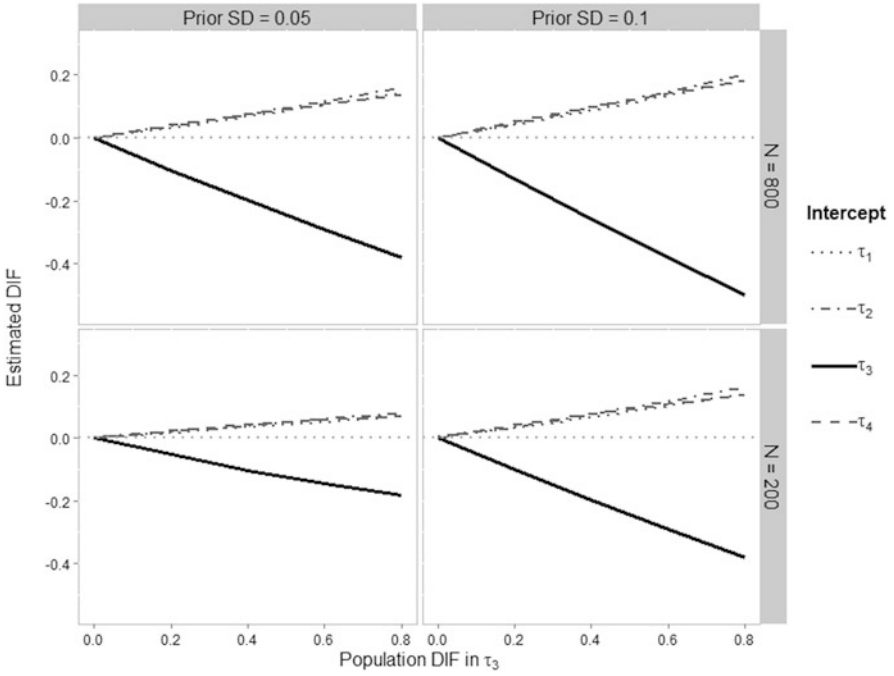


Fig. 3 Average posterior mean of $\Delta\tau$s by DIF, prior σ, and *N*, with separate lines per $\Delta\tau$

1 and 2(b and f) represented exact rather than approximate metric and (full or partial) scalar invariance.

The practical impact of these biased estimates can be reflected by rates at which the H_0 of invariance was rejected. Figure 4 compares Type I error rates (averaged across non-DIF parameters) between ML modification indices (grey lines) and 95% BCIs (black lines). While Type I error rates fluctuated around the nominal 5% for modification indices, the BCIs had near-zero error rates across conditions. As typically happens when Type I error rates are higher, power for modification indices was also somewhat higher in some conditions (see Fig. 5).

4 Discussion

The use of parameter estimates constrained by small-variance priors as a Bayesian analog to ML modification indices (Muthén & Asparouhov, 2012) seems to have some limited potential. Their power to detect DIF is often similar to (sometimes lower than) modification indices, but they have lower Type I error rates. However, small-variance priors continue to propagate bias throughout the model, just as invalid exactly-zero constraints do in ML estimation. So it may not be advisable

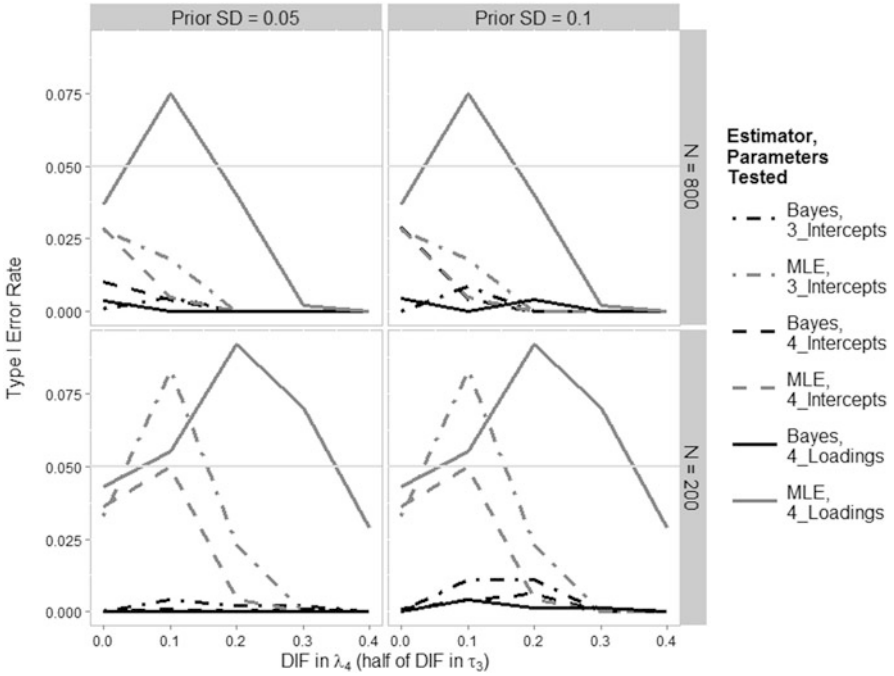


Fig. 4 Type I error rates by DIF, prior σ, and *N*, with separate lines per estimator and model

to use small-variance priors, even following a sensitivity analysis to choose their precision (e.g., Asparouhov et al., 2015). At the very least, relying on parameter estimates constrained by small-variance priors for clues about necessary model modifications (which Muthén & Asparouhov, 2012, indicated was a "side product of the proposed approach", p. 313) does not imply that models with small-variance priors should be used for inference.

As Muthén and Asparouhov (2012) assert in their subtitle, small-variance priors for nontarget parameters are intended to provide researchers with a more flexible representation of [their] substantive theory. But because PPP appears insensitive to minor misspecification (Jorgensen et al., 2019), nontarget parameters could potentially be fixed to zero without PPP indicating poor model fit. When misspecification is too severe to be ignorable, PPP would have even greater power to reject the model if priors for nontarget parameters were excluded altogether (i.e., nontarget parameters fixed to zero). When a SEM without small-variance priors indicates poor (exact or even approximate) fit, small-variance priors for nontarget parameters could then be added to help detect the local source of misfit; however, the priors should be only weakly informative to increase the probability that they indicate a neglected parameter should be "freed" and (contrary to Muthén and Asparouhov's advice) freed one parameter at a time rather than considering all parameters simultaneously. Future research should explore the possibility of

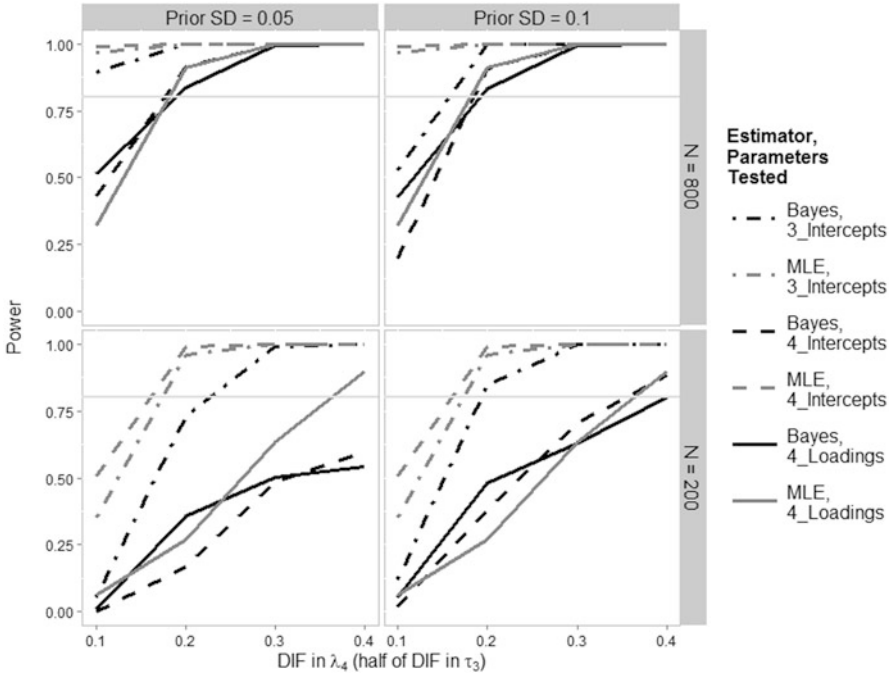


Fig. 5 Power by DIF, prior σ, and *N*, with separate lines per estimator and model

developing more reliable tools to detect local sources of misspecification in BSEM (i.e., sensitive to misspecification without propagating errors throughout the model), perhaps using a PPMC framework to investigate score-based statistics, analogous to actual modification indices in ML estimation.

References

Asparouhov, T., Muthén, B., & Morin, A. J. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeyer et al. *Journal of Management, 41*(6), 1561–1577. https://doi.org/10.1177/0149206315591075

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Garnier-Villarreal, M., & Jorgensen, T. D. (2020). Adapting fit indices for Bayesian structural equation modeling: Comparison to maximum likelihood. *Psychological Methods, 25*(1), 46–70. https://doi.org/10.1037/met0000224

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*, 733–807. https://doi.org/10.1.1.142.9951

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. *Sociological Methods & Research, 26*, 329–367. https://doi.org/10.1177/0049124198026003003

Jorgensen, T. D. (2015). *Selecting an optimal measurement model and detecting differential item functioning using Bayesian confirmatory factor analysis* [Doctoral dissertation, University of Kansas]. https://doi.org/10.13140/RG.2.2.14104.03841.

Jorgensen, T. D., Garnier-Villarreal, M., Pornprasertmanit, S., & Lee, J. (2019). Small-variance priors can prevent detecting important misspecifications in Bayesian confirmatory factor analysis. In M. Wiberg, S. A. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology: The 83rd annual meeting of the psychometric society, New York, 2018* (pp. 255–263). Springer. https://doi.org/10.1007/978-3-030-01310-3_23

Levy, R. (2011). Bayesian data–model fit assessment for structural equation modeling. *Structural Equation Modeling, 18*(4), 663–685. https://doi.org/10.1080/10705511.2011.607723

Merkle, E. C., & Rosseel, Y. (2018). `Blavaan`: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software, 85*(4), 1–30. https://doi.org/10.18637/jss.v085.i04

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313–335. https://doi.org/10.1037/a0026802

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Author.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Proper and Useful Distractors in Multiple-Choice Diagnostic Classification Models

Hans Friedrich Köhn, Chia-Yi Chiu, and Yu Wang

Abstract The multiple-choice (MC) item format has been implemented in educational assessments that are used across diverse content domains. MC items comprise two components: the stem that provides the context with a motivating narrative, and the collection of response options consisting of the correct answer, called the "key," and several incorrect alternatives, the "distractors." The MC-DINA model was the first diagnostic classification model for MC items that used distractors explicitly as potential sources of diagnostic information. However, the MC-DINA model requires that the q-vectors of the distractors are nested within each other and that of the key, which poses a serious constraint on item development. Consequently, later adaptations of the MC item format to cognitive diagnosis dropped the nestedness condition. The relaxation of the nestedness-condition, however, comes at a price: distractors may become redundant (i.e., they do not contribute to any further diagnostic differentiation between examinees), and they may induce undesirable diagnostic ambiguity (i.e., they are equally likely to be chosen by an examinee, but their q-vectors point at different diagnostic classifications). In this article, two criteria, *useful* and *proper*, are proposed to identify redundant and diagnostically ambiguous distractors.

Keywords Cognitive diagnosis · Nonparametric cognitive diagnosis · Polytomous items · MC-DINA · MC-NPC

H. F. Köhn (✉)
Department of Psychology, University of Illinois, Urbana-Champaign, IL, USA
e-mail: hkoehn@illinois.edu

C.-Y. Chiu · Y. Wang
Educational Psychology, University of Minnesota Twin Cities, Minneapolis, MN, USA
e-mail: cchiu@umn.edu; wang7919@umn.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_9

1 Introduction

The multiple-choice (MC) item format has been implemented in educational assessments that are used across diverse content domains. MC items comprise two components: the stem prepares examinees for the test questions in providing the context and a motivating narrative; the collection of response options contains the correct answer, called the "key," and several incorrect alternatives, the "distractors." Different from the dichotomous response format, MC items allegedly permit for collecting richer diagnostic information, while examinees need to spend less time on recording their answers. MC items are also less vulnerable to subjective scoring. In summary, the economy of MC items is likely one of the reasons for their persistent popularity not just in educational testing.

The MC item format has been adapted to accommodate also the cognitive diagnosis (CD) framework in educational measurement. Within CD, ability—or competence—in a curricular knowledge domain is perceived as a composite of cognitive skills called "attributes." CD-based tests consist of items that require for a correct response mastery of different attributes. From the item responses, examinees' ability can be inferred and evaluated in terms of attributes mastered and those needing study.

Early approaches to analyzing MC items within the CD framework lacked sophistication such that the MC responses were simply dichotomized in scoring the key as 1 and the distractors as 0, (e.g., Lee et al., 2011; Templin & Henson, 2006). The dichotomized responses were then analyzed using one of the CD models—diagnostic classification model (DCM) hereafter—for binary responses. De la Torre's (2009) Multiple-Choice Deterministic Inputs, Noisy "And" Gate (MC-DINA) model was the first DCM for analyzing MC items in considering the distractors explicitly as potential sources of diagnostic information. Recall that in case of the DINA model, an item can only discriminate between examinees, who have mastered all required attributes and those who fail one or more of these attributes. In contrast, extending the model to accommodate the MC item format, is expected to increase the classification accuracy in allowing for the separation of examinees into more than two groups. To this purpose, de la Torre's (2009) MC-DINA model, as a particular feature, requires that the attribute profiles of the distractors be nested within the attribute profile of the key. But obviously, such a nestedness requirement puts an undue burden on the test developer, as the options for item building are seriously constrained. In fact, later adaptations of the MC item format to CD—including the current implementation of de la Torre's (2009) MC-DINA in the `R` package `GDINA` (Ma & de la Torre, 2020)—have abandoned the nestedness requirement (e.g., Ozaki, 2015; DiBello et al., 2015; Wang et al., 2021).

The relaxation of the nestedness-condition, however, comes at a price. First, distractors may become redundant; that is, they do not contribute to any further diagnostic differentiation between examinees. But redundancy among distractors may not be easy to detect; in addition, the inclusion of redundant distractors in

a test results in wasting valuable item space and may increase unsystematic error variance. Second, distractors may create undesirable diagnostic ambiguity; that is, they are equally likely to be chosen by an examinee, but their attribute profiles point at different diagnostic classifications.

In this article, a rationale is developed based on psychometric theory to detect these two cases during test construction before the test is actually used in the field. Specifically, we propose two criteria that identify *useful* and *proper* distractors. A distractor is said to be *useful* if it is not redundant. A redundant distractor is one that does not improve the classification of examinees beyond the response options already available for a given item. A distractor is called *proper* if it allows for the unambiguous identification of an examinee's ideal response. Notice that in the case of the MC-DINA model the restriction that all distractors must be nested within each other prevents such potential ambiguity.

The next section briefly reviews essential CD concepts and their adaptation to accommodate MC items like the MC-DINA model and the nonparametric classification method for MC items (MC-NPC). Section 3 presents the theory of useful and proper distractors. The discussion section concludes with a summary of the key insights and a discussion of some limitations and future research avenues.

2 Review of Key Technical Concepts

2.1 Cognitive Diagnosis

DCMs for cognitive diagnosis (CD), a formative assessment framework in educational measurement, describe ability in a given knowledge domain as a composite of K cognitive skills—henceforth: "attributes"—that a student has mastered or not (DiBello et al., 2007; Haberman & von Davier, 2007; Leighton & Gierl, 2007; Nichols et al., 1995; Rupp et al., 2010; Sessoms & Henson, 2018; Tatsuoka, 2009). Attribute mastery is recorded as a K-dimensional binary vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k, \ldots, \alpha_K)' \in \{0, 1\}^K$. Distinct attribute profiles $\boldsymbol{\alpha}_m$ identify different classes of proficiency $\mathcal{C}_m$, $m = 1, 2, \ldots, M = 2^K$ (provided the attributes are not hierarchically organized). (The terms profile and vector are used interchangeably here.) The primary task of CD is to assign students to one of these M classes based on their performance in a test that targets proficiency in the knowledge domain in question. Said differently, examinees' individual attribute vectors $\boldsymbol{\alpha}_{i \in \mathcal{C}_m}$ must be estimated ($i = 1, \ldots, N$ is the examinee index; for brevity, $\boldsymbol{\alpha}_{i \in \mathcal{C}_m} = \boldsymbol{\alpha}_m = \boldsymbol{\alpha}_i$ is often used, depending on the context).

CD items require mastery of domain-specific attributes for a correct response. Similar to examinees, CD items are characterized by individual K-dimensional attribute profiles $\mathbf{q}_j$, with entries $q_{jk} = 1$ if a correct answer requires mastery of the k^{th} attribute α_k, and 0 otherwise ($j = 1, 2, \ldots, J$ is the item index). (Notice that

the zero vector is not admissible; thus, there are at most $2^K - 1$ distinct item-attribute profiles.)

The collection of the item attribute profiles of a CD assessment forms its Q-matrix $\mathbf{Q} = \{q_{jk}\}_{(J \times K)}$ (Tatsuoka, 1982) that establishes the associations between items and attributes. The Q-matrix of a test must be known and it must be complete. A Q-matrix is said to be complete if its specific composition can guarantee the identifiability of all realizable proficiency classes among examinees (Chiu et al., 2009; Köhn & Chiu, 2016; 2017; 2019; 2021). Q-completeness is the key requirement for the identifiability of DCMs.

2.2 The MC-DINA Model

The intuitive appeal of the simple conceptual elegance of the Deterministic Inputs, Noisy "AND" Gate (DINA) model (Haertel, 1989; Junker & Sijtsma, 2001; Macready & Dayton, 1977) is presumably the reason why the DINA model is arguably one of the most popular DCMs. The DINA is a conjunctive model, as the probability of a correct response is maximal only if an examinee has mastered all attributes required for a given item. Thus, each DINA item generates a bi-partition of the $M = 2^K$ proficiency classes of the latent attribute space into groups of examinees who have mastered the attributes required for said item as opposed to those who have not. The item response function (IRF) of the DINA model is

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i) = (1 - s_j)^{\eta_{ij}} g_j^{(1-\eta_{ij})} \tag{1}$$

where $Y_{ij} = 1$ denotes the correct response to item j (otherwise, $Y_{ij} = 0$); s_j and g_j are slipping and guessing parameters, respectively, subject to $0 \leq g_j < 1 - s_j \leq 1$; the ideal response (or conjunction parameter) $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} = 0, 1$ indicates whether examinee i has mastered all attributes required by item j.

The MC-DINA model was proposed by de la Torre (2009) as an extension of the DINA model to accommodate MC items. Recall that each DINA item results in a bi-partition of the latent attribute space. In contrast, an MC item partitions the latent attribute space into a number of proficiency classes that is proportional to those of coded response options, thereby purportedly increasing the accuracy of examinee classification. (A response option is said to be "coded" or "cognitively based" if it is linked to an item attribute vector $\mathbf{q}_{jh}$ specifying the attribute requirements for an examinee who endorses this option; terminology and notation follow de la Torre, 2009.) In case of the MC-DINA model, the polytomous response to item j is denoted as the random variable X_{ij}, with the response options indexed by $h = 1, 2 \ldots, H_j$. Let H_j^* denote the number of coded options. Since not all options are coded, $H_j^* \leq H_j$. "Non-coded" response options like "none of these" or "all of the above" are not associated with a specific attribute vector. Hence, as a convention, their item attribute vectors are written as a K-dimensional null

vector, $\mathbf{q}_{j0} = (0, 0, \ldots, 0)'$. The key has always the largest number of attributes; the attribute vectors of the coded response options must be nested within the q-vector of the key. Their attribute vectors must also be hierarchically nested within each other such that they form an ordinal scale with the key at the top.

2.3 Removing the Nestedness Condition

If the q-vectors of the coded response options are not required to be nested within each other and the q-vector of the key, then the linear ordering of the coded response options is lost and their scale level is changed to nominal. As an adjustment to this significant conceptual modification, the original MC-DINA response option index, $h = 1, 2, \ldots, H_j$, is replaced by the index $l = 0, 1, 2, \ldots, H_j^*$, and the notation for the q-vector $\mathbf{q}_{jh}$ is changed to one involving the index l: $\mathbf{q}_j^{(l)}$. All non-coded response options are indexed as $l = 0$, having q-vectors $\mathbf{q}_j^{(0)}$. The key is indexed as $l = H_j^*$; thus, $\mathbf{q}_j^{(H_j^*)}$. The indices $l = 1, 2, \ldots, H_j^* - 1$, are assigned to the remaining coded response options according to the following rationale.

Let $\mathbf{q}$ and $\mathbf{q}'$ denote the q-vectors of distinct coded response options. If $\| \mathbf{q}_j \|_1 > \| \mathbf{q}'_j \|_1$, then $l > l'$ so that the notation becomes $\mathbf{q}_j^{(l)}$ and $\mathbf{q}_j^{(l')}$. (and vice versa; $\| \cdot \|_1$ denotes the L_1 norm). If $\mathbf{q}$ and $\mathbf{q}'$ are of the same length, then the response options are indexed based on their evaluation in lexicographic order—that is, $l > l'$ if the position of the first non-zero entry in $\mathbf{q}$ precedes that in $\mathbf{q}'$, and vice versa. Ties—both q-vectors share the position of the first non-zero entry—are ignored and the evaluation is based on the first position with distinct entries; such a position can always be identified because all coded response option q-vectors must be distinct. Formally, define the set $\mathcal{L}(\mathbf{q}, \mathbf{q}') = \{k \mid q_k > q'_k, k = 1, 2, \ldots, K\}$. Notice that $\mathcal{L}(\mathbf{q}, \mathbf{q}') \neq \mathcal{L}(\mathbf{q}', \mathbf{q})$ due to the evaluation of q_k and q'_k in lexicographic order. If

$$\left(\| \mathbf{q}_j^{(l)} \|_1 = \| \mathbf{q}_j^{(l')} \|_1 \right) \wedge \left(\min \mathcal{L}\left(\mathbf{q}_j^{(l)}, \mathbf{q}_j^{(l')}\right) < \min \mathcal{L}\left(\mathbf{q}_j^{(l')}, \mathbf{q}_j^{(l)}\right) \right)$$

then $l > l'$.

After the indices l of the item response options have been determined, the ideal response η_{ij} of examinee i to item j can be computed

$$\eta_{ij} = \max_{l=0,1,2,\ldots,H_j^*} \left\{ l \prod_{k=1}^{K} I[\alpha_{ik} \geq q_{jk}^{(l)}] \right\} \tag{2}$$

where $I[\cdot]$ denotes the indicator function.

The original IRF of the MC-DINA model is not provided in de la Torre (2009); however, for the case that the nestedness condition has been removed, the IRF of the MC-DINA model can be (re-)constructed as

$$P(X_{ij}=l \mid \boldsymbol{\alpha}_i)=\begin{cases}\frac{1}{H_j}+\frac{H_j-H_j^*-1}{H_j}\, I[l=0] & \text{if } \eta_{ij}=0\\ \left(1-\sum_{m\neq l}\varepsilon_{jml}\right)^{I[\eta_{ij}=l,l>0]}\prod_{l'\neq l}\varepsilon_{jll'}^{I[\eta_{ij}=l']} & \text{if } \eta_{ij}>0\end{cases} \tag{3}$$

where $\varepsilon_{jll'}$ is the probability that the observed response level l disagrees with the ideal response level l'. (Because the manifest and ideal item responses, X and η, now have more than two levels, addressing potential discrepancies between X and η as "slips" and "guesses" does not fit the complexity of the MC setting involving multiple item parameters. The more general term "perturbation" should be preferred whenever observed and ideal responses disagree.) Typically, slipping and guessing are constrained to be less than 0.5 (otherwise, an individual mastering none of the attributes would have a probability greater than 0.5 to provide the correct answer). Of course, if there are more than two perturbation terms, then the desirable property is that $\sum_{m\neq l}\varepsilon_{jml} < 0.5$. An examinee with $\eta_{ij}=0$ is not "attracted" (de la Torre, 2009) to any of the coded response options. Instead, said examinee is assumed to pick one of the response options at random. If $(\eta_{ij}=l)\wedge(l\neq 0)$, then examinee i is supposed to choose the coded response option $X_{ij}=l$ with high probability; still, alternative response options may be chosen with non-zero probability.

2.4 The MC-NPC Method

The MC-NPC method is the nonparametric counterpart to the MC-DINA model. As was mentioned earlier, different from de la Torre's (2009) original MC-DINA, for the MC-NPC, like for the MC-DINA implementation in the `R` package `GDINA`, the q-vectors of the coded distractors are not required to be nested within each other and the q-vector of the key. Thus, the item response options have nominal scale level.

MC-NPC is an adaptation of the nonparametric classification (NPC) method (Chiu & Douglas, 2013) to accommodate the MC item format for the DINA model. "Nonparametric" refers to the fact that the NPC methods do not rely on the parametric estimation of examinees' proficiency class membership, but use a distance-based algorithm on the observed item responses for classifying examinees. Proficiency class membership is determined by comparing an examinee's observed item response vector $\mathbf{X}$ with each of the ideal item response vectors of the M realizable proficiency classes. Let $\boldsymbol{\eta}^{(m)}=\left(\eta_1^{(m)},\eta_2^{(m)},\ldots,\eta_J^{(m)}\right)$ denote the ideal item response vector of examinees in $\mathcal{C}_m$ as defined in Eq. (2). An examinee's proficiency class is identified by the attribute vector $\boldsymbol{\alpha}_m$ underlying that ideal item response vector which, among all ideal response vectors, minimizes the penalized Hamming distance to the manifest item response vector $\mathbf{X}_i=\mathbf{x}_i$. The penalized Hamming distance is defined as

Table 1 Example to illustrate non-useful distractors

(a)

Option	Level	q
Key	4	(100)
Distractor	3	(111)
Distractor	2	(110)
Distractor	1	(001)

(b)

$\boldsymbol{\alpha}$	(000)	(100)	(010)	(001)	(110)	(101)	(011)	(111)
η	0	4	0	1	4	4	1	4

$$d_p(\mathbf{x}_i, \boldsymbol{\eta}^{(m)}) = \sum_{j=1}^{J} I[\eta_j^{(m)} > 0, X_{ij} \neq \eta_j^{(m)}] + \sum_{j=1}^{J} w_j I[\eta_j^{(m)} = 0, X_{ij} \neq \eta_j^{(m)}] \tag{4}$$

The estimate of the attribute profile of examinee i is identified by minimizing $d_p(\mathbf{x}_i, \boldsymbol{\eta}^{(m)})$ across all ideal response profiles $\boldsymbol{\eta}^{(1)}, \boldsymbol{\eta}^{(2)}, \ldots, \boldsymbol{\eta}^{(M)}$ and observed response profile $\mathbf{X}_i = \mathbf{x}_i$:

$$\widehat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_m \in \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_M\}}{\arg\min} \; d_p(\mathbf{x}_i, \boldsymbol{\eta}^{(m)})$$

3 Coded Response Options: The Concepts of Proper and Useful

Recall that the nestedness condition imposed on the key and coded response options may create significant difficulties when constructing a test simply because the number of such nested coded response options may be limited—hence, the proposition to replace the nestedness condition by the two more flexible criteria of *useful* and *proper* distractors.

A distractor is said to be *useful* if it is not redundant. A redundant distractor is one that does not improve the classification of examinees beyond the response options already available for a given item. Here is an illustration of the concept. Suppose the q-vector of the key of an item is (100) and those of the three distractors are (111), (110), and (001). These four response options are coded as 4, 3, 2, and 1, respectively, as shown in Table 1(a).

For the 2^K possible attribute profiles, the corresponding ideal responses are computed using Eq. (2) and listed in Table 1(b). Notice that the levels of the response options cover the entire range from 0 to 4; however, only three different ideal responses, 0, 1, and 4, can be identified based on the information provided by the key and the coded distractors. Hence, distractors having levels 2 and 3 are redundant

because they do not provide any information to identify their corresponding ideal responses.

A formal definition of a "useful" distractor is provided below, but some notation is needed first. "Nestedness" is denoted by "$\prec$." Specifically, for vectors $\mathbf{a} = (a_1, \ldots, a_K)$ and $\mathbf{b} = (b_1, \ldots, b_K)$, $\mathbf{a} \prec \mathbf{b}$ if and only if $a_k \leq b_k$ for all k but $\mathbf{a} \neq \mathbf{b}$. Also, $\mathbf{a} \preceq \mathbf{b}$ if and only if $a_k \leq b_k$ for all k. Let $\mathcal{L}$ be the latent space of all realizable attribute profiles. Define the set $\mathcal{L}_j = \{\boldsymbol{\alpha} \mid \boldsymbol{\alpha} \not\succeq \mathbf{q}_{jH_j^*}\}$ consisting of all attribute profiles that *do not* contain all the attributes required by the key of item j—said differently, all attribute profiles that do not allow an examinee to answer item j correctly. Define $\mathcal{G}_j(\boldsymbol{\alpha}) = \left\{h \mid \boldsymbol{\alpha}^\top \mathbf{q}_h = \mathbf{q}_h^\top \mathbf{q}_h, \text{where } h \in \{0, \cdots, H_j^* - 1\}, \boldsymbol{\alpha} \in \mathcal{L}_j\right\}$ as the set of the indices h of distractors that are *nested* within $\boldsymbol{\alpha}$. If $\mid \mathcal{G}_j(\boldsymbol{\alpha}) \mid > 1$, then define $\mathcal{G}_j^*(\boldsymbol{\alpha}) = \{h \mid \mathbf{q}_h \not\prec \mathbf{q}_{h'}, \text{where } h, h' \in \mathcal{G}_j\}$ as the subset of $\mathcal{G}_j(\boldsymbol{\alpha})$ that contains the indices h of all response options that are *not nested* within those in $\mathcal{G}_j(\boldsymbol{\alpha})$. Notice that if $\mathcal{G}_j(\boldsymbol{\alpha})$ contains only a single element, then $\mathcal{G}_j^*(\boldsymbol{\alpha}) = \mathcal{G}_j(\boldsymbol{\alpha})$.

A distractor h, with $h \in \{1, \cdots, H_j^* - 1\}$, is defined as *useful* if there exists an $\boldsymbol{\alpha} \in \mathcal{L}_j$ such that $h \in \mathcal{G}_j^*(\boldsymbol{\alpha})$. The condition that identifies a *useful* distractor is summarized in the following claim: a distractor h of item j is *useful* if and only if $\mathbf{q}_{jH_j^*} \not\preceq \mathbf{q}_{jh}$.

A distractor is called *proper* if it allows for the nonambiguous identification of an examinee's ideal response. Notice that in case of the MC-DINA model the restriction that all distractors must be nested within each other prevents such ambiguity. As an example, consider an item where the key is coded as (1111) and the two distractors as (1100) and (0110). Notice that they are not nested within each other; hence, they may induce ambiguity about an examinee's classification. For $\boldsymbol{\alpha} = (1110)$, the ideal response is $\eta(1110) = \max\limits_{l \in \{0,1,2,3\}} \left\{l \prod_{k=1}^{4} I[\alpha_{ik} \geq q_{jk}^{(l)}]\right\} = 2$ (see Eq. (2)). So, an examinee having $\boldsymbol{\alpha} = (1110)$ is expected to choose Distractor 2. However, because $(1110) \succ (1100)$ and $(1110) \succ (0110)$, this examinee is equally likely to choose Distractor 1. This potential mismatch between the ideal and observed response induces ambiguity concerning an examinee's proficiency class and may cause her misclassification.

To avoid such ambiguity, the concept of *proper* distractors is introduced by the following definition: the distractors of item j are said to be *proper* if $\mid \mathcal{G}_j^*(\boldsymbol{\alpha}) \mid = 1$ for all $\boldsymbol{\alpha} \in \mathcal{L}_j$. Now, let $Q_j = \{\mathbf{q}_{j1}, \cdots, \mathbf{q}_{jH_j^*}\}$ be the set consisting of the q-vectors of all the coded options for item j. In addition, suppose the distractors of item j are *useful*. Then they are claimed to be *proper* if and only if for each pair of coded distractors with $\mathbf{q}_{jh}$ and $\mathbf{q}_{jh'}$, where $h, h' \in \{1, \cdots, H_j^* - 1\}$, and $\bigcup(\mathbf{q}_{jh}, \mathbf{q}_{jh'}) \neq \left(1, 1, \ldots, 1\right)'$, then $\bigcup(\mathbf{q}_{jh}, \mathbf{q}_{jh'}) \in Q_j$.

4 Discussion and Outlook

The concepts of *useful* and *proper* distractors of MC items in CD assessments, as they are presented in this article explore uncharted territory. Redundancy and

diagnostic ambiguity never were an issue, as long as the nestedness condition implied tight control over the diagnostic utility of the distractors of MC items. However, the relaxation of the nestedness constraint in recent approaches to CD modeling of MC items has created redundancy and diagnostic ambiguity as new challenges in item construction for researchers as well as educational practitioners.

Avenues for future research concern (1) a broader theoretical appraisal of *useful* and *proper* as criteria for evaluating the quality of distractors of MC items, and (2) simulation studies to assess the extent to which distractors failing these two criteria induce examinee misclassification.

References

Chiu, C.-Y., & Douglas, J. A. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response profiles. *Journal of Classification, 30*, 225–250.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*, 633–665.

de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement, 33*, 163–183.

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 26: Psychometrics* (pp. 979–1030). Elsevier.

DiBello, L. V., Henson, R. A., & Stout, W. F. (2015). A family of generalized diagnostic classification models for multiple choice option-based scoring, *Applied Psychological Measurement, 39*, 62–79.

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 26: Psychometrics* (pp. 1031–1038). Elsevier.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333–352.

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

Köhn, H.-F., & Chiu, C.-Y. (2016). Conditions of completeness of the Q-matrix of tests for cognitive diagnosis. In L. A. van der Ark, D. M. Bolt, W. C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative Psychology Research: The 80th Annual Meeting of the Psychometric Society* (pp. 255–264). Springer.

Köhn, H.-F., & Chiu, C.-Y. (2017). A procedure for assessing completeness of the Q-matrices of cognitively diagnostic tests. *Psychometrika, 82*, 112–132.

Köhn, H.-F., & Chiu, C.-Y. (2019). Attribute hierarchy models in cognitive diagnosis: Identifiability of the latent attribute space and conditions for completeness of the Q-matrix. *Journal of Classification, 36*, 541–565.

Köhn, H.-F., & Chiu, C.-Y. (2021). A unified theory of the completeness of Q-matrices for the DINA model. *Journal of Classification, 38*(3), 500–518.

Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing, 11*(2), 144–177.

Leighton, J., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software, 93*(14), 1–26.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 33*, 379–416.

Nichols, P. D., Chipman, S. E., & Brennan, R. L. (1995). *Cognitively diagnostic assessment*. Erlbaum.

Ozaki, K. (2015). DINA models for multiple-choice items with few parameters: Considering incorrect answers. *Applied Psychological Measurement, 39*, 431–447.

Rupp, A. A., & Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford.

Sessoms, J., & Henson, R. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives, 16*, 1–17.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconception in the pattern classification approach. *Journal of Educational and Behavioral Statistics, 12*, 55–73.

Tatsuoka, K. K. (2009). *Cognitive assessment. An introduction to the rule space method*. Routledge/Taylor & Francis.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305.

Wang, Y., Chiu, C.-Y., & Köhn, H. F. (2021). Nonparametric classification method for multiple-choice items in cognitively diagnostic assessments. In *Paper presented at the virtual annual meeting of the Psychometric Society*.

Detecting Latent Variable Non-normality Through the Generalized Hausman Test

Lucia Guastadisegni, Irini Moustaki, Vassilis Vasdekis, and Silvia Cagnone

Abstract This paper extends the generalized Hausman test to detect non-normality of the latent variable distribution in unidimensional IRT models for binary data. To build the test, we consider the estimator obtained from the two-parameter IRT model, that assumes normality of the latent variable, and the estimator obtained under a semi-nonparametric framework, that allows for a more flexible latent variable distribution. The behaviour of the test is evaluated through a simulation study. The results highlight the good performance of the test in terms of both Type I error rates and power with many items and large sample sizes.

Keywords Generalized Hausman test · SNP-IRT model · Binary data

1 Introduction

One of the typical assumptions of latent variable models is the normal distribution of the latent variables. As shown in Ma and Genton (2010), this assumption is not always appropriate and misspecifying the form of the latent variable by assuming normality can result in large biases in parameter estimates. Several methods, that assume a different form for the latent variable, have been proposed within the generalized latent variable models (GLLVM) and Item Response Theory (IRT) framework. Some examples are the semi-parametric (Ma & Genton, 2010), the empirical histogram (Knott & Tzamourani, 2007), the Ramsey-curve (Woods, 2006)

L. Guastadisegni (✉) · S. Cagnone
University of Bologna, Bologna, Italy
e-mail: lucia.guastadisegni2@unibo.it; silvia.cagnone@unibo.it

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

V. Vasdekis
Athens University of Economics and Business, Athens, Greece
e-mail: vasdekis@aueb.gr

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_10

and the semi-nonparametric (SNP) (Gallant & Nychka, 1987, Woods & Lin, 2009, Irincheeva et al., 2012) methods.

Commonly information criteria are used to choose between a model where the latent variables are normal and a model where they have a more complex shape (Woods & Lin, 2009, Irincheeva et al., 2012). However, detecting non-normality of the latent variables through a statistical test remains an open issue.

Hausman (1978) proposes a specification test to detect failure of the orthogonality assumption in the regression model. The Hausman test can be applied also in other contexts, to detect different types of model misspecification. The idea of the test is simple. It compares two different estimators that are consistent when the model is correctly specified and one is also efficient. In presence of model misspecification, only the inefficient estimator is consistent. The efficiency assumption simplifies the computation of the covariance matrix of the difference between the two estimators. However, this matrix can fail to be positive definite under model misspecification or in presence of small sample sizes. A generalized version of the Hausman (GH) test has been proposed by White (1982). In this case none of the estimators that result from different models need to be efficient and the covariance matrix involved in the test is always positive definite.

As far as we know, in the IRT context the classic Hausman test has been used only by Ranger and Much (2020) to detect misspecification of the item characteristic functions and local dependencies among items. In generalized linear mixed models (GLMM) for clustered data, a robust version of the Hausman test, similar to the one by White (1982), has been proposed by Bartolucci et al. (2017) when a discrete distribution for the random effects is assumed.

The objective of this work is to extend the GH test to detect non-normality of the latent variable distribution in unidimensional IRT models for binary data. To build the test, we consider the estimators resulting from two different models and estimation methods. The first model is the classical unidimensional IRT model for binary data based on the normality assumption of the latent variable, where we estimate the parameters using a maximum pairwise likelihood (PL) method. The PL method uses information from bivariate-order margins and belongs to the family of composite likelihood methods (Lindsay, 1988, Varin, 2008). It produces biased parameter estimates when the latent variable is not normally distributed. The second model is the unidimensional SNP-IRT model for binary data (Woods & Lin, 2009, Irincheeva et al., 2012), and we estimate the parameters using the quasi-maximum likelihood (ML) method. The choice of these estimators for the two models is motivated by the following reasons. First, both methods are consistent when the latent variable is normally distributed. Moreover, the quasi-ML method for the SNP_L model is consistent also under different distribution assumptions of the latent variable (Gallant & Tauchen, 1989, Irincheeva et al., 2012). These conditions on the consistency of the parameter estimators are required to correctly apply the Generalized Hausman test (White, 1982). Second, the maximum PL estimator is less efficient than the ML estimator. This implies that, also under normality of the latent variable distribution, the covariance matrix of the difference of the two estimators

involved in the GH test is different from zero. This allows us to avoid numerical problems in the computation of the test.

The article is organized as follows. First, we review the classical and SNP-IRT model for binary data. Second, we introduce the GH test to detect non-normality of the latent variable distribution. Next, we present a Monte Carlo simulation study. Finally, we present some concluding remarks.

2 The Classical and SNP-IRT Model for Binary Data

Let us denote by $y_1, \ldots, y_p$ a set of observed binary variables/items, by n the number of individuals and by z the latent variable with density function $h(z)$.

For the classical IRT model, the response category probability for the i-th individual to the j-th item is modelled using a logistic model (measurement model)

$$P(y_{ij} = 1|z_i) = \pi_{ij}(z_i) = \frac{\exp(\alpha_{0j} + \alpha_{1j} z_i)}{1 + \exp(\alpha_{0j} + \alpha_{1j} z_i)}, \tag{1}$$

where α_{0j} is the item intercept and α_{1j} the item slope. In this model $h(z) = \phi(z)$, where $\phi(z)$ is the density of a standard normal.

For the SNP-IRT model, the response probability is the same as (1), where the latent variable has a SNP parametrization

$$h(z_i) = P_L^2(z_i)\phi(z_i) \qquad P_L(z_i) = \sum_{0 \le l \le L} a_i z_i^l, \tag{2}$$

$a_0, \ldots, a_L$ are the real coefficients of the polynomial $P_L(z_i)$ and L is the polynomial degree.

In order for $h(z)$ to be a density, the coefficients $a_0, \ldots, a_L$ of $P_L(z)$ should be chosen such that $\int h(z)dz = 1$. For this purpose, Gallant and Tauchen (1989) use a proportionality constant $1/\int P_L(z)^2\phi(z)dz$ and fix the constant term of the polynomial equal to 1. Alternatively, Irincheeva et al. (2012) and Woods and Lin (2009) use the parametrization proposed by Zhang and Davidian (2001), that imposes

$$1 = \int_R P_L^2(z)\phi(z)dz = E\{P_L^2(w)\} = a' E(\tilde{w}\tilde{w}')a = a' A a \tag{3}$$

with $w \sim N(0, 1)$, $P_L(w) = a'\tilde{w}$, $\tilde{w} = (1, w, w^2, \ldots, w^L)$. The matrix A is positive definite by definition and $A = B'B$, where B is a positive definite matrix.

If $c = Ba$, Eq. (3) becomes $c'c = 1$ and $c = (c_1, \ldots, c_{L+1})'$. The elements of c can be represented using a polar coordinate transformation as $c_1 = \sin\varphi_1, c_2 = \cos\varphi_1 \sin\varphi_2, \ldots, c_L = \cos\varphi_1 \times \cos\varphi_{L-1} \sin\varphi_L, c_{L+1} = \cos\varphi_1 \cos\varphi_2 \times \cos\varphi_{L-1} \cos\varphi_L$, with angles $-\pi/2 < \varphi_t \le \pi/2, t = 1, \ldots, L$. The

density of the latent variable in (2) can be expressed as

$$h(z|\boldsymbol{\varphi}, L) = (a'\tilde{\mathbf{z}})^2\phi(z), \tag{4}$$

where a can be obtained from c as $a = B^{-1}c$, $\tilde{\mathbf{z}} = (1, z, z^2, \ldots, z^L)'$ and $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_L)'$.

When $L = 1$, $P_L(z) = a_0 + a_1 z$, $a_0 = \sin\varphi_1$, $a_1 = \cos\varphi_1$. When $L = 0$ the distribution of the latent variable reduces to the normal one. In the following sections we indicate with SNP_1 the model for $L = 1$ and with SNP_0 the model for $L = 0$.

2.1 Pairwise Estimator for the SNP_0 Model

To implement the GH test, the parameters of the SNP_0 model are estimated with the pairwise method. The pairwise log-likelihood of the data, based on the bivariate marginal densities $f(y_{ij}, y_{ik}, \boldsymbol{\theta})$, $j, k = 1, ..p$ and $k > j$, is

$$\begin{aligned} pl_{SNP_0}(\mathbf{y}, \boldsymbol{\theta}) &= \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k>j} \ln f(y_{ij}, y_{ik}, \boldsymbol{\theta}) = \\ &= \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k>j} \ln \int \left[\pi_{ij}(z_i)^{y_{ij}}(1-\pi_{ij}(z_i))^{1-y_{ij}}\right] \\ &\quad \times \left[\pi_{ik}(z_i)^{y_{ik}}(1-\pi_{ik}(z_i))^{1-y_{ik}}\right]\phi(z_i)dz_i. \end{aligned} \tag{5}$$

The pairwise log-likelihood is maximized with respect to $\boldsymbol{\theta}$, that includes the item intercepts and slopes. Under correct model specification, the maximum PL estimator $\tilde{\boldsymbol{\theta}}$ converges in probability to the true parameter value $\boldsymbol{\theta}_0$ and

$$\tilde{\boldsymbol{\theta}} \xrightarrow{p} N(\boldsymbol{\theta}_0, A^{-1}(\boldsymbol{\theta}_0)B(\boldsymbol{\theta}_0)A^{-1}(\boldsymbol{\theta}_0)), \tag{6}$$

where $A(\boldsymbol{\theta}) = E_y\left[-\frac{\partial^2 pl_{SNP_0}(\mathbf{y},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]$, $B = var\left[\frac{\partial pl_{SNP_0}(\mathbf{y},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]$ and $A(\boldsymbol{\theta}) \neq B(\boldsymbol{\theta})$ (Lindsay, 1988, Varin, 2008). These matrices can be estimated by their observed versions as

$$\hat{A}(\boldsymbol{\theta}) = -\sum_{i=1}^{n}\frac{\partial^2 pl_{SNP_0}(\mathbf{y}_i, \boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \tag{7}$$

and

$$\hat{B}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial pl_{SNP_0}(\boldsymbol{y}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial pl_{SNP_0}(\boldsymbol{y}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}. \tag{8}$$

2.2 Quasi-ML Estimator for the SNP_L Model

The parameters of the SNP_L model, $L > 0$, are estimated with the quasi-ML method. The log-likelihood of the data is

$$\begin{aligned} l_{SNP_L}(\mathbf{y}, \boldsymbol{\theta}) &= \sum_{i=1}^{n} \ln f(\mathbf{y}_i, \boldsymbol{\theta}) = \\ &= \sum_{i=1}^{n} \ln \int \prod_{j=1}^{p} \pi_{ij}(z_i)^{y_{ij}} (1 - \pi_{ij}(z_i))^{1-y_{ij}} P_L^2(z_i) \exp\left(-\frac{1}{2} z_i' z_i \right) dz_i. \end{aligned} \tag{9}$$

The integral in the log-likelihood $l(\mathbf{y}, \boldsymbol{\theta})$ is approximated with the Gauss-Hermite quadrature, as in Woods and Lin (2009). The degree of the polynomial L is fixed and is not estimated by maximum likelihood. The log-likelihood function is maximized with respect to the unknown vector of parameter $\boldsymbol{\theta} = (\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \boldsymbol{\varphi})$ as follows

$$(\dot{\boldsymbol{\alpha}}_0, \dot{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\varphi}}) = argmax_{\boldsymbol{\theta}} l_{SNP_L}(\mathbf{y}, \boldsymbol{\theta}). \tag{10}$$

For identifiability reasons, the item intercepts and slopes, that correspond to a latent variable that has mean 0 and variance 1, are rescaled as (Irincheeva et al., 2012)

$$\hat{\alpha}_{0j} = \dot{\alpha_0}_j + \dot{\alpha_1}_j \tilde{E}(Z) \qquad j = 1, \ldots, p \tag{11}$$

$$\hat{\alpha}_{1j} = \dot{\alpha_1}_j \sqrt{\tilde{V}(Z)} \qquad j = 1, \ldots, p, \tag{12}$$

where $\tilde{E}(Z)$ and $\tilde{V}(Z)$ are found given $\hat{\boldsymbol{\varphi}}$ and the SNP density of z. The final quasi-ML estimator is $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\varphi}})$. Under normal, multi-modal and asymmetric distributions of the latent variables and if the regularity conditions A2–A6 of White (1982) are satisfied,

$$\hat{\boldsymbol{\theta}} \xrightarrow{p} N(\boldsymbol{\theta}_{0*}, A^{-1}(\boldsymbol{\theta}_{0*}) B(\boldsymbol{\theta}_{0*}) A^{-1}(\boldsymbol{\theta}_{0*})), \tag{13}$$

where $\boldsymbol{\theta}_{0*}' = (\boldsymbol{\alpha}_{00}', \boldsymbol{\alpha}_{01}', \boldsymbol{\varphi}_*')$. $\boldsymbol{\alpha}_{00}$ and $\boldsymbol{\alpha}_{01}$ are the true parameter values for the item intercepts and slopes while $\boldsymbol{\varphi}_*$ is the value of $\boldsymbol{\varphi}$ that minimizes the Kullback-Leibler information criterion (White, 1982, Gallant and Tauchen, 1989, Irincheeva et al., 2012). $A(\boldsymbol{\theta})$ and $B(\boldsymbol{\theta})$ are the expected Hessian and cross-product matrices, respectively. Their observed versions can be computed with the Delta method

(Cramér, 1946) and are defined similarly to (7) and (8), where $pl_{SNP_0}(\mathbf{y}_i, \boldsymbol{\theta})$ is replaced by $l_{SNP_L}(\mathbf{y}_i, \boldsymbol{\theta})$.

3 The Generalized Hausman Test

In this section we present the GH test, derived by White (1982), applied to detect non-normality of the latent variable using the SNP-IRT model.

Let's denote by $\boldsymbol{\eta}$ the sub-vector of $\boldsymbol{\theta}' = (\boldsymbol{\alpha}_0', \boldsymbol{\alpha}_1', \boldsymbol{\varphi}')$ that includes the item intercepts $\boldsymbol{\alpha}_0$ and slopes $\boldsymbol{\alpha}_1$. $\boldsymbol{\eta}$ has dimension $2p \times 1$, where p is the number of items.

Consider the maximum PL estimator $\tilde{\boldsymbol{\theta}}_{SNP_0} = \tilde{\boldsymbol{\eta}}_{SNP_0}$ of a classic IRT model where the latent variable is normally distributed, that is the SNP_0 model.

Consider the quasi-ML estimator $\hat{\boldsymbol{\theta}}'_{SNP_L} = (\hat{\boldsymbol{\eta}}'_{SNP_L}, \hat{\boldsymbol{\varphi}}')$ of a SNP-IRT model with $L > 0$, where the sub-vector of parameter $\hat{\boldsymbol{\varphi}}$ has dimension $L \times 1$ and so $\hat{\boldsymbol{\theta}}_{SNP_L}$ has dimension $(2p + L) \times 1$. Following White (1982), under normality of the latent variable

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{SNP_L} - \tilde{\boldsymbol{\eta}}_{SNP_0}) \xrightarrow{d} N(0, S(\boldsymbol{\eta}_0, \boldsymbol{\theta}_{0*})). \tag{14}$$

An estimator of $S(\boldsymbol{\eta}_0, \boldsymbol{\theta}_{0*})$ is

$$\begin{aligned}\hat{S}(\tilde{\boldsymbol{\eta}}_{SNP_0}, \hat{\boldsymbol{\theta}}_{SNP_L}) = {} & \hat{A}^{\boldsymbol{\eta}\boldsymbol{\varphi}}(\hat{\boldsymbol{\theta}}_{SNP_L})^{-1} \hat{B}(\hat{\boldsymbol{\theta}}_{SNP_L}) \hat{A}^{\boldsymbol{\eta}\boldsymbol{\varphi}}(\hat{\boldsymbol{\theta}}_{SNP_L})^{-1'} \\ & + \hat{A}(\tilde{\boldsymbol{\eta}}_{SNP_0})^{-1} \hat{B}(\tilde{\boldsymbol{\eta}}_{SNP_0}) \hat{A}(\tilde{\boldsymbol{\eta}}_{SNP_0})^{-1'} \\ & - \hat{A}^{\boldsymbol{\eta}\boldsymbol{\varphi}}(\hat{\boldsymbol{\theta}}_{SNP_L})^{-1} \hat{R}(\tilde{\boldsymbol{\eta}}_{SNP_0}, \hat{\boldsymbol{\theta}}_{SNP_L})' \hat{A}(\tilde{\boldsymbol{\eta}}_{SNP_0})^{-1'} \\ & - \hat{A}(\tilde{\boldsymbol{\eta}}_{SNP_0})^{-1} \hat{R}(\tilde{\boldsymbol{\eta}}_{SNP_0}, \hat{\boldsymbol{\theta}}_{SNP_L}) \hat{A}^{\boldsymbol{\eta}\boldsymbol{\varphi}}(\hat{\boldsymbol{\theta}}_{SNP_L})^{-1'},\end{aligned} \tag{15}$$

where the matrices $\hat{A}(\tilde{\boldsymbol{\eta}}_{SNP_0})$ and $\hat{B}(\tilde{\boldsymbol{\eta}}_{SNP_0})$, defined in formulas (7) and (8), have dimension $2p \times 2p$ and are evaluated at $\tilde{\boldsymbol{\eta}}_{SNP_0}$. $\hat{A}(\hat{\boldsymbol{\theta}}_{SNP_L})$ and $\hat{B}(\hat{\boldsymbol{\theta}}_{SNP_L})$ are the observed Hessian and cross-product matrix of dimension $(2p + L) \times (2p + L)$ for the SNP_L model, evaluated at $\hat{\boldsymbol{\theta}}_{SNP_L}$. The matrix $\hat{A}^{\boldsymbol{\eta}\boldsymbol{\varphi}}(\hat{\boldsymbol{\theta}}_{SNP_L})^{-1}$ is obtained by deleting the last L row from the matrix $\hat{A}(\hat{\boldsymbol{\theta}}_{SNP_L})^{-1}$ and has dimension $2p \times (2p + L)$. The matrix $\hat{R}(\tilde{\boldsymbol{\eta}}_{SNP_0}, \hat{\boldsymbol{\theta}}_{SNP_L})$ has dimension $2p \times (2p + L)$ and can be computed as

$$\hat{R}(\boldsymbol{\eta}_{SNP_0}, \boldsymbol{\theta}_{SNP_L}) = \sum_{i=1}^{n} \frac{\partial pl_{SNP_0}(\boldsymbol{y}_i, \boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \frac{\partial l_{SNP_L}(\boldsymbol{y}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}, \tag{16}$$

where $pl_{SNP_0}(\boldsymbol{y}_i, \boldsymbol{\eta})$ is the pairwise log-likelihood for the individual i under the model SNP_0 and $l_{SNP_L}(\boldsymbol{y}_i, \boldsymbol{\theta})$ is the log-likelihood for the individual i under the model SNP_L. The matrix in (16) is evaluated at $(\tilde{\boldsymbol{\eta}}_{SNP_0}, \hat{\boldsymbol{\theta}}_{SNP_L})$. We choose the

maximum PL and the quasi-ML estimator for the two models to avoid that, under correct model specification, $\tilde{\boldsymbol{\eta}}_{SNP_0}$ and $\hat{\boldsymbol{\eta}}_{SNP_L}$ converge to the same covariance matrix, producing a $\hat{S}(\tilde{\boldsymbol{\eta}}_{SNP_0}, \hat{\boldsymbol{\theta}}_{SNP_L})$ matrix in (15) with all entries close to 0.

Given the theoretical result in (14), the GH test is given by

$$GH = (\hat{\boldsymbol{\eta}}_{SNP_L} - \tilde{\boldsymbol{\eta}}_{SNP_0})' \hat{S}(\tilde{\boldsymbol{\eta}}_{SNP_0}, \hat{\boldsymbol{\theta}}_{SNP_L})^{-1} (\hat{\boldsymbol{\eta}}_{SNP_L} - \tilde{\boldsymbol{\eta}}_{SNP_0}). \tag{17}$$

Under normality of the latent variable, the GH test is asymptotically distributed as a χ^2_{2p}, where $2p$ are the degrees of freedom, i.e. the number of parameters in $\boldsymbol{\eta}$.

However, the matrix $\hat{S}(\tilde{\boldsymbol{\eta}}_{SNP_0}, \hat{\boldsymbol{\theta}}_{SNP_L})$ is often close to singularity and its inversion in formula (17) is numerically unstable.

Given the theoretical result in (14) and the quadratic form $(\hat{\boldsymbol{\eta}}_{SNP_L} - \tilde{\boldsymbol{\eta}}_{SNP_0})'(\hat{\boldsymbol{\eta}}_{SNP_L} - \tilde{\boldsymbol{\eta}}_{SNP_0})$, we consider the following test statistic (Ranger & Much, 2020)

$$GH_T = (\hat{\boldsymbol{\eta}}_{SNP_L} - \tilde{\boldsymbol{\eta}}_{SNP_0})'(\hat{\boldsymbol{\eta}}_{SNP_L} - \tilde{\boldsymbol{\eta}}_{SNP_0}). \tag{18}$$

Under normality of the latent variable

$$GH_T \sim \sum_{l=1}^{d} \lambda_l z_l^2, \qquad z_l \sim N(0, 1), \tag{19}$$

where d is the rank of $S(\boldsymbol{\eta}_0, \boldsymbol{\theta}_{0*})$ and $\lambda_1, \ldots, \lambda_d$ are its non-zero eigenvalues.

It is possible to approximate the distribution in (19) as follows (Welch, 1938,Yuan & Bentler, 2010)

$$GH_T \sim a\chi^2_b. \tag{20}$$

The quantity a and b are defined as

$$a = \frac{\sum_{l=1}^{d} \lambda_l^2}{\sum_{l=1}^{d} \lambda_l} \tag{21}$$

and

$$b = \frac{(\sum_{l=1}^{d} \lambda_l)^2}{\sum_{l=1}^{d} \lambda_l^2}. \tag{22}$$

Since $S(\boldsymbol{\eta}_0, \boldsymbol{\theta}_{0*})$ can be consistently estimated by $\hat{S}(\tilde{\boldsymbol{\eta}}_{SNP_0}, \hat{\boldsymbol{\theta}}_{SNP_L})$ defined in (15), a and b can be consistently estimated substituting $\hat{\lambda}_1, \ldots, \hat{\lambda}_d$ in (21) and (22), where d is rank of $\hat{S}$ and $\hat{\lambda}_1, \ldots, \hat{\lambda}_d$ are its non-zero eigenvalues.

4 Simulation Study

4.1 Simulation Design

In this section we study the performance of the GH_T test by a simulation study. The estimation of the SNP-IRT model is computationally expensive. Moreover, as the degree of the polynomial L increases ($L > 1$), the SNP_L model becomes more sensitive to the choice of the initial values for all model parameters and the estimation results can be less reliable. Furthermore, in the data generating models we assume the latent variable distributed as mixtures of two normals, that can be well approximated with $L = 1$, as highlighted in Irincheeva et al. (2012). Thus, to implement the GH_T test, we consider the SNP_0 and the SNP_1 models. The optimization of the SNP_0 model is obtained with direct maximization using the function "optim" of the software R while, for the SNP_1 model, the function "nlminb", that makes use of the analytically computed gradient and Hessian matrix. For the SNP_1 model, initial values of the parameters $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ are the parameter estimates obtained with the SNP_0 model. In each data replication, for the φ_1 parameter, we sample 10 initial values from a sequence of values equally spaced by 0.1 in the interval $[-\frac{\pi}{2}; \frac{\pi}{2}]$, i.e. the domain of φ_1, including the SNP_0 model as a subcase. Among the estimated SNP_1 models in each data replication, we select the one that corresponds to the maximum value of the log-likelihood function. All matrices involved in the GH_T test are computed numerically with the "NumDeriv" R package. Although assuming a SNP distribution for the latent variable is more computationally demanding than assuming the normal distribution, it has the great advantage that it is very flexible and produces accurate estimates in many situations.

We consider the following simulation conditions: number of items ($p = 4, 10, 20$) $\times$ sample size ($n = 500, 1000$) $\times$ test statistic (GH_T). In all the simulation scenarios, $R = 500$ replications are considered and $\alpha = 0.05$. Non-valid statistics, for example negative statistics, are excluded from the analysis. The Type I error rates and power of the GH_T test are computed as $\hat{p} = \sum_{l=1}^{N_v} \frac{I(GH_{T_l} \geq c)}{N_v}$, where N_v is the number of valid statistics out of the number of replications, I is an indicator function, GH_{T_l} is the value of the GH_T test statistic evaluated in the l-th replication. c is the theoretical asymptotic critical value corresponding to the $(1-\alpha)$th percentile of the $a\chi_b^2$ distribution for the GH_T test, where a and b are computed as in (21) and (22). The confidence interval (CI) of each rate $\hat{p}$ is computed as $\hat{p} \pm z_{(1-\frac{\alpha}{2})}\sqrt{\frac{\alpha(1-\alpha)}{N_v}}$.

To evaluate the performance of the GH_T test, we consider three scenarios (SC), corresponding to three different distribution assumptions for the latent variable z in the data generating models. The general model is

$$\begin{aligned} &logit(\pi_{ij}) = \alpha_{0j} + \alpha_{1j} z_i \qquad i = 1, \ldots, n \qquad j = 1, 2, \ldots, p \\ &z \sim h(z) \end{aligned} \tag{23}$$

Item intercepts are randomly generated in the interval $[-0.8; 1.12]$ while the item slopes in the interval $[0.5; 1.5]$.

To study the Type I error rates of the GH_T test we consider the following scenario:

A $z \sim N(0, 1)$

To study the power of the GH_T test we consider the following two scenarios:

B $z \sim 0.1N(-2, 0.25) + 0.9N(2, 1)$,
where z has an overall mean equal to 1.6 and variance equal to 2.365.

C $z \sim 0.7N(-1.5, 0.6) + 0.3N(1.5, 0.5)$,
where z has an overall mean equal to -0.6 and variance equal to 2.217.

Under the distributional assumptions of the two scenarios B and C, the estimates of the quasi-ML parameters of the SNP_1 model are nearly unbiased (see the results on the bias of the parameters in scenario B reported in Irincheeva et al., 2012) while the maximum PL parameter estimates of the SNP_0 model are largely biased with respect to the true parameter values. This should result in a good GH_T test performance in terms of power.

4.2 Results

Table 1 reports the Type I error rates, mean and standard deviation of the theoretical(T) and empirical(E) distribution of the GH_T test for scenario A.

Overall, the GH_T test has good performance in terms of Type I error rates when the sample size is large and in general with many items. Moreover, the empirical

Table 1 Type I error rates, mean and standard deviation of the theoretical(T) and empirical(E) distribution of the GH_T test for scenario A, $p = 4, 10, 20$, $n = 500, 1000$

p	n	Distribution	Mean	SD	α
4	500	TD	2.01	2.00	0.050
		ED	1.61	1.81	**0.016**
	1000	TD	2.12	2.06	0.050
		ED	2.54	2.93	**0.086**
10	500	TD	3.44	2.62	0.050
		ED	2.89	2.64	**0.018**
	1000	TD	3.21	2.54	0.050
		ED	3.00	2.97	0.044
20	500	TD	3.48	2.64	0.050
		ED	3.44	3.15	0.056
	1000	TD	3.52	2.65	0.050
		ED	3.63	3.12	0.060

Note 1: Values in boldface indicate that the nominal level α is not included in their confidence interval

Table 2 Empirical power of the GH_T test for scenarios B and C, $p = 4, 10, 20$, $n = 500, 1000$

SC	p	n	Power
B	4	500	0.53
		1000	0.86
	10	500	0.924
		1000	0.998
	20	500	0.99
		1000	0.998
C	4	500	0.796
		1000	0.92
	10	500	1
		1000	1
	20	500	0.986
		1000	1

distribution of the GH_T test approaches the theoretical one as the number of items and the sample size increase. Small differences can be found in terms of empirical and theoretical standard deviations, while the means of the two distributions are very similar under most conditions. Despite the good performance of the test with many items and large sample size in terms of Type I error rates, we observe an inconsistent pattern of results with 4 items and all sample sizes. In general, the estimation of the model parameters and the related information matrices, on which the GH_T test is based, is less accurate on small data sets. Indeed, few items and small sample sizes carry out less information than more items and large sample sizes. We should consider larger sample sizes to obtain Type I error rates of the GH_T test close to the nominal level α for 4 items, while $n = 1000$ is sufficient for 10 items and $n = 500$ for 20 items.

Table 2 presents the power of the GH_T test for scenarios B and C.

The power of the GH_T test is high when the sample size is large and with 10 and 20 items. Moreover, it increases with the number of items and the sample size.

5 Conclusion

In this work, we extended the GH test to detect non-normality of the latent variable distribution in unidimensional IRT models for binary data. The GH test was obtained as the difference between the estimators of the classic IRT model for binary data and the SNP-IRT model, that allows for a more flexible shape of the latent variable distribution. To avoid the inversion of the covariance matrix of the difference between the parameter estimates, we considered an alternative form of this test, that we called GH_T test, and we evaluated its performance by means of a small simulation study.

The simulation study highlights that the GH_T test has good performance in terms of Type I error rates with many items and in particular for large sample sizes. For what concerns the power, the GH_T test has good performance with many items and large sample sizes. However, these are preliminary results. Further studies should include other distributions of the latent variables. Indeed, it would be interesting to study the behaviour of this test when the SNP approach performs less well in recovering the distribution of the latent variable, for example when it is skewed (Monroe, 2014). Moreover, the GH_T test presented in this work could be applied to IRT models for polytomous data, assuming the SNP representation of the latent variable distribution. Since these models involve a higher number of parameters, the additional issue, compared to binary data, could be the computational cost of the estimation process (Bartholomew et al., 2011).

The GH test could also be applied to detect other types of model violations, as local dependence or violation of the item characteristic function. In these cases, other types of estimators consistent under model misspecification should be considered in order to apply the test.

References

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (3rd ed.). Wiley.

Bartolucci, F., Bacci, S., & Pigini, C. (2017). Misspecification test for random effects in generalized linear finite-mixture models for clustered binary and ordered data. *Econometrics and Statistics, 3*, 112–131.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton University Press.

Gallant, A. R., & Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica, 55*(2), 363–390.

Gallant, A. R., & Tauchen, G. (1989). Seminonparametric estimation of conditionally constrained heterogeneous processes: Asset pricing applications. *Econometrica, 57*(5), 1091–1120.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica, 46*(6), 1251–1271.

Irincheeva, I., Cantoni, E., & Genton, M. G. (2012). Generalized linear latent variable models with flexible distribution of latent variables. *Scandinavian Journal of Statistics, 39*(4), 663–680.

Knott, M., & Tzamourani, P. (2007). Bootstrapping the estimated latent distribution of the two-parameter latent trait model. *British Journal of Mathematical and Statistical Psychology, 60*(1), 175–191.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics, 80*(1), 221–239.

Ma, Y., & Genton, M. G. (2010). Explicit estimating equations for semiparametric generalized linear latent variable models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72*(4), 475–495.

Monroe, S. L. (2014). *Multidimensional item factor analysis with semi-nonparametric latent densities*. Ph.D Thesis, UCLA.

Ranger, J., & Much, S. (2020). Analyzing the fit of IRT models with the Hausman test. *Frontiers in Psychology, 11*, 149.

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis, 92*(1), 1–28.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29*(3–4), 350–362.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50*(1), 1–25.

Woods, C. M. (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods, 11*(3), 253–270.

Woods, C. M., & Lin, N. (2009). Item response theory with estimation of the latent density using Davidian curves. *Applied Psychological Measurement, 33*(2), 102–117.

Yuan, K.-H., & Bentler, P. M. (2010). Two simple approximations to the distributions of quadratic forms. *British Journal of Mathematical and Statistical Psychology, 63*(2), 273–291.

Zhang, D., & Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics, 57*(3), 795–802.

A Speed-Accuracy Response Model with Conditional Dependence Between Items

Peter W. van Rijn and Usama S. Ali

Abstract Conditional independence assumptions play an important role in many psychometric models, but can sometimes be too restrictive in modeling process data from educational and psychological tests such as response times. For this reason, a continuous speed-accuracy response model is developed that relaxes the assumption of conditional independence of items given latent proficiency ("local" independence). Our model is a generalization of the speed-accuracy response model developed by Maris and van der Maas (Psychometrika, 77:615-633, 2012) in which a scoring rule incorporating both accuracy and speed of item responses is assumed to produce a sufficient statistic for a latent proficiency variable. The assumption of local independence is dropped in a similar way as in the interaction model developed for dichotomous item responses by Haberman (Multivariate and Mixture Distribution Rasch Models, pp. 201–216. Springer, New York, 2007). Recently, Verhelst (Theoretical and Practical Advances in Computer-Based Educational Measurement, pp. 135–160. Springer, Cham, 2019) discussed similar models in the context of exponential family models for continuous item responses. A pairwise conditional maximum likelihood approach is developed to estimate item parameters. The model is illustrated by an application to data from a listening test.

Keywords Conditional independence · Interaction model · Speed-accuracy response models · Pairwise conditional maximum likelihood

P. W. van Rijn (✉)
ETS Global, Amsterdam, The Netherlands
e-mail: pvanrijn@etsglobal.org

U. S. Ali
Educational Testing Service, Princeton, NJ, USA

South Valley University, Qena Governorate, Egypt
e-mail: uali@ets.org; usama.ali@edu.svu.edu.eg

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_11

1 Introduction

Conditional independence assumptions play an important role in the way psychometric models are applied. For example, the assumption of "local" independence in item response theory (IRT) models (Lord & Novick, 1968), which states that item responses are independent conditional on a latent proficiency variable, greatly simplifies the application of these models to real data. In general, such assumptions can be problematic when modeling process data from educational and psychological tests such as response times to test items (Bolsinova et al., 2017). Problems with these assumptions can arise both within items (e.g., between accuracy and speed) and between items (e.g., due to speededness). Our objective is to develop and estimate a speed-accuracy response model (Maris & van der Maas, 2012) which permits conditional dependencies between items similar to the interaction model for discrete item responses (Haberman, 2007). This would allow a more flexible way of dealing with response-time data, but also with other continuous item response data (Verhelst, 2019).

We start the next section with models for dichotomous item responses. This is followed by a section on speed-accuracy response models, in which we present our new model. Next, we discuss model estimation using a pairwise likelihood approach and an application to real data. The paper ends with a brief discussion.

2 Discrete Item-Response Models

Under local independence, the log probability of the vector $\mathbf{y}$ containing m dichotomous item responses under the Rasch (1960) model can be given by

$$\log p(\mathbf{y}|\theta) = C(\theta) + \sum_{j=1}^{m} y_j(\theta + \beta_j), \tag{1}$$

where $C(\theta)$ is a normalizing factor, θ is a latent proficiency variable, and β_j is the intercept parameter for item j. The normalizing factor $C(\theta)$ for the Rasch model is $-\sum_{j=1}^{m} \log\left[1 + \exp(\theta + \beta_j)\right]$. An important feature under the Rasch model is that the total score $r = \sum_{j=1}^{m} y_j$ is a sufficient statistic for θ (Andersen, 1970).

Haberman's (2007) interaction model is an extension of the Rasch model with conditional dependence between items:

$$\log p(\mathbf{y}|\theta) = C(\theta) + \sum_{j=1}^{m} y_j(\theta + \beta_j) + \sum_{j=2}^{m}\sum_{k=1}^{j-1} y_j y_k(\gamma_j + \gamma_k), \tag{2}$$

$$= C(\theta) + r\theta + \boldsymbol{\beta}'\mathbf{y} + (r-1)\boldsymbol{\gamma}'\mathbf{y}. \tag{3}$$

where γ_j is the interaction parameter for item j. It can be seen from Eq. (3) that item difficulty is effectively a linear function of the total score. In this model, the total score remains a sufficient statistic for θ, but local independence is no longer assumed.

Conditional on the total score r, we obtain the following conditional probability for the interaction model

$$\log p(\mathbf{y}|r) = D(r) + \boldsymbol{\beta}'\mathbf{y} + (r-1)\boldsymbol{\gamma}'\mathbf{y}, \tag{4}$$

where the conditional normalizing factor $D(r)$ is

$$D(r) = -\log \sum_{\mathbf{y}:r} \exp\left[\boldsymbol{\beta}'\mathbf{y} + (r-1)\boldsymbol{\gamma}'\mathbf{y}\right]. \tag{5}$$

The summation here runs over all response pattern $\mathbf{y}$ that result in total score r. Equation (4) is independent of θ and can be used to develop conditional maximum likelihood (CML) estimation for item parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. CML estimates of item parameters of both the Rasch and Haberman model can be obtained with the R package dexter (Maris et al., 2022).

3 Speed-Accuracy Response Models

In discussing speed-accuracy response models, we focus on the signed residual time (SRT) scoring rule:

$$x_j = (2y_j - 1)(d_j - t_j), \tag{6}$$

where x_j is the score awarded to the speed and accuracy of the response to item j, d_j is a time limit for item j and $t_j > 0$ is the continuous response time for item j. In this scoring rule, fast correct responses are given more credit than slow correct responses, and fast incorrect responses are penalized more than slow incorrect responses. Of course, there are many more ways to combine speed and accuracy into a single value, each with merits and limitations, but we limit our attention to the SRT score here. Note that response times can be rescaled so that d_j can be fixed to one and $-1 < x_j < 1$ for all items.

The speed-accuracy response model from Maris and van der Maas (2012) is:

$$\log f(\mathbf{x}|\theta) = C(\theta) + \sum_{j=1}^{m} x_j(\theta + \beta_j). \tag{7}$$

The Maris–van der Maas model is related to the Rasch model for continuous item responses (Müller, 1987), but also has marginal functions for item responses and

response times. For example, the marginal function for item responses, or the item response function, turns out to be the two-parameter logistic (2PL) model where the item discrimination equals the time limit d_j (Maris & van der Maas, 2012, Eq. (12)). The speed-accuracy response model of Eq. (7) and an extension including a discrimination parameter can be estimated using marginal maximum likelihood (MML) estimation, for example, with the dedicated software package SARM (van Rijn & Ali, 2018b).

Our proposed new model is the Maris–van der Maas model extended with conditional dependence between items:

$$\log f(\mathbf{x}|\theta) = C(\theta) + \sum_{j=1}^{m} x_j(\theta + \beta_j) + \sum_{j=2}^{m}\sum_{k=1}^{j-1} x_j x_k(\gamma_j + \gamma_k). \tag{8}$$

As in the Haberman interaction model, the total score remains a sufficient statistic for θ, but local independence does not hold. The Haberman model provides an approximation to a 2PL model and our new model does so in a similar fashion to the two-parameter speed-accuracy response model by van Rijn and Ali (2018b). However, no distributional assumption is needed for θ to estimate Haberman's and our new model.

The normalizing factor $C(\theta)$ is difficult to obtain for our model, because the scores are continuous. A similar issue also occurs in other models with interactions (e.g., the partition function in Boltzmann distributions, Ising models; Maris & Bechger, 2021). However, things prove to be easier when we focus on item pairs (j, k) with $1 \leq k < j \leq m$ instead of the full vector and condition on their total score $r = x_j + x_k$, $j \neq k$.

Conditional on $r_{jk} = x_j + x_k$, we can then write

$$\log f(x_j, x_k|r) = D(r_{jk}) + x_j\beta_j + x_k\beta_k + x_j x_k(\gamma_j + \gamma_k). \tag{9}$$

The pairwise conditional normalizing factor $D(r)_{jk}$ can be obtained and turns out to be

$$D(r_{jk}) = -\log \int_{\tilde{r}_{jkl}}^{\tilde{r}_{jku}} \exp[x\beta_j + (r_{jk} - x)\beta_k + x(r_{jk} - x)(\gamma_j + \gamma_k)]dx, \tag{10}$$

where $\tilde{r}_{jkl} = \max(r_{jk} - 1, -1)$ and $\tilde{r}_{jku} = \min(r_{jk} + 1, 1)$. The integral can be solved analytically using the error function, but this can give numerical issues. Instead, numerical integration (adaptive quadrature) can be used. We now have an expression (Eq. (9)) for the conditional probability of item pairs which is independent of θ. Note that if a scoring rule different from the SRT is used (e.g., the unsigned residual time $Y_j(d_j - T_j)$), the normalizing factors would need to be derived accordingly.

4 Estimation

A pseudo-likelihood approach can now be developed to estimate item parameters $\boldsymbol{\xi} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$ (Besag, 1975). Specifically, we use the pairwise conditional log likelihood (PCL; Verhelst, 2019)

$$PCL(\boldsymbol{\xi}) = \sum_{i=1}^{n} \sum_{j=2}^{m} \sum_{k=1}^{j-1} \log f(x_{ij}, x_{ik} | r_{ijk}) \tag{11}$$

$$= \sum_{i=1}^{n} \sum_{j=2}^{m} \sum_{k=1}^{j-1} D(r_{ijk}) + \sum_{j=2}^{m} \sum_{k=1}^{j-1} [s_j \beta_j + s_k \beta_k + s_{jk}(\gamma_j + \gamma_k)], \tag{12}$$

where x_{ij} is the SRT score of test taker i on item j, r_{ijk} is the total score on item pair (j, k) for test taker i, $s_j = \sum_{i=1}^{n} x_{ij}$ and $s_{jk} = \sum_{i=1}^{n} x_{ij} x_{ik}$. To identify the model, we fix $\beta_m = \gamma_m = 0$. The main challenge lies in the computation of the conditional normalizing factor $D(r_{ijk})$ for which we use numerical integration. The PCL can be maximized numerically or otherwise.

If the interaction terms are dropped, the PCL can also be used for estimating $\boldsymbol{\beta}$ in the Maris–van der Maas model. Note that in this case only so-called weak local independence is assumed (i.e, item pairs are independent conditional on θ; McDonald, 1999). The PCL approach also works for the Rasch model (Zwinderman, 1995), but not for the Haberman model. This can be illustrated, for example, using the probability of getting item j correct and item k incorrect, given that either j or k is correct (i.e., the sum score is 1). Under the Haberman model, this probability simplifies to $p(1, 0|1) = \frac{\exp(\beta_j)}{\exp(\beta_j)+\exp(\beta_k)}$, which is the same as in the Rasch model. This means that we cannot use pairwise probabilities to distinguish the Rasch and the Haberman model. The use of triplewise probabilities could put us back in business. However, when the number of items becomes more than 5 there are increasingly more triples than pairs to deal with, making the approach impractical for longer tests. Such triplewise conditional likelihood estimation would nevertheless be of interest (e.g., in case of missing or incomplete data), but is beyond our present scope.

Various aspects of pairwise CML (PCML) estimation such as efficiency and uniqueness could be studied further, but space is limited for our current exposition. An important aspect though is that there are dependencies when using pairwise probabilities (van der Linden & Eggen, 1986, p. 347). For example, if a person responds to three dichotomously scored items a, b, c, and $y_a = 0$ and $y_b = 1$ is observed, then $y_a = 1$ and $y_c = 0$ cannot be observed. The same holds in a more intricate fashion for continuous scores (e.g., if $x_a > x_b$, then $x_a < x_c$ is less likely). These dependencies limit the use of the pairwise likelihood (e.g., for model comparisons).

5 Application

To illustrate the new model, we make use of 17 items from a listening section of an English language test for non-native speakers. The sample size is $n = 9355$. Since the data were not collected under item-specific time limits (although there was a time limit on the section), we leniently created the time limits d_j for the SRT scoring rule based on the 99-th percentile of the empirical response time distributions (as was done in van Rijn & Ali, 2018a). We estimate the Rasch, Haberman, Maris–van der Maas, and our new model. We compared the CML and PCML estimates for the Rasch model, and the MML and PCML estimates for the Maris–van der Maas model as a check to see if the pairwise estimation works. Furthermore, the CML estimates for the Haberman model and the PCML estimates of our new model are compared as well.

Figure 1 shows the results of the check on pairwise estimation. CML estimates of item parameters for the Rasch and Haberman model were obtained with the R package dexter (Maris et al., 2022). MML estimates for the Maris–van der Maas model were obtained with the SARM software (van Rijn & Ali, 2018b). The PCML estimates for both models were obtained with our own R code. As can be seen from the figure, PCML produces very similar estimates as CML for the Rasch model and as MML for the Maris–van der Maas model. Note that some care regarding model identification is needed when comparing estimates across CML and MML estimation.

Figure 2 displays a comparison of the $\boldsymbol{\beta}$ parameters between the Maris–van der Maas and our new model and a comparison of the $\boldsymbol{\gamma}$ parameters between the Haberman and our new model. The estimated $\boldsymbol{\beta}$s of the Maris–van der Maas and our new model are very similar. The estimated $\boldsymbol{\gamma}$s for the Haberman and our new model are correlated, but on different scale. This difference in scale is not surprising

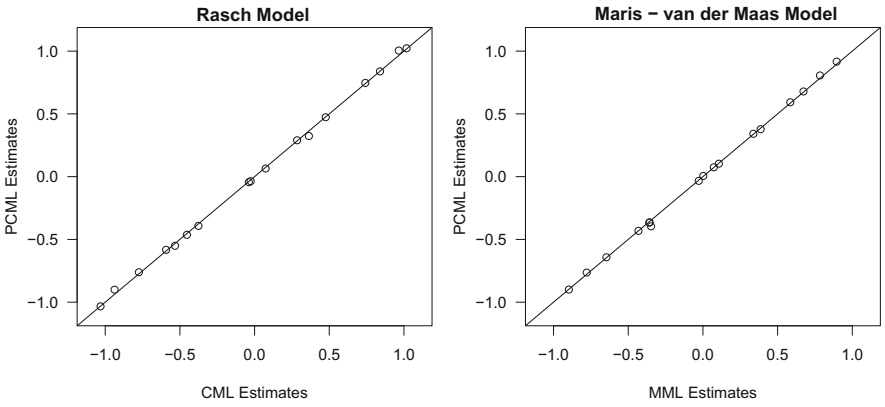


Fig. 1 Comparison of CML and PCML estimates of item parameters for Rasch model (left) and MML and PCML for Maris–van der Maas model (right) for listening test

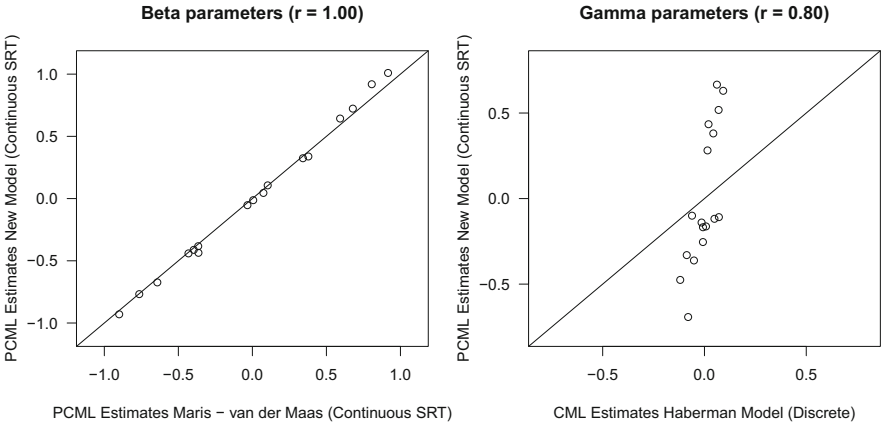


Fig. 2 Comparison of item parameters across Maris–van der Maas and new model (left) and across Haberman and new model (right) for listening test

given that these models are fitted to different data (dichotomous item responses vs. continuous scores based on the SRT scoring rule).

6 Discussion

In this paper, we presented a new speed-accuracy response model with conditional dependence between items. We developed a PCML estimation procedure to estimate its item parameters, which, on the basis of the application, appears to work, so that our new model can be practically estimated. Nevertheless, various estimation aspects need to be sorted out (e.g., efficiency, uniqueness, dependencies among pairwise probabilities, information loss). In addition, whether our model fits better than simpler models remains to be determined.

A limitation of our exposition is that we only focus on the SRT scoring rule, which is not widely used in practice. However, a benefit of our model and estimation procedure is that it can also be applied to other types of continuous scores, potentially including process data. We argue that this is relevant since it is likely that item responses will be scored more and more frequently by algorithms than by humans, and such algorithms often produce a continuous score. It would be interesting also to investigate whether a mix of different scores can be modelled within this approach.

Another benefit of the new model is that no distributional assumption is needed for θ to estimate its structural parameters. This can be important especially when the quality of the sample of test takers is diverse. In addition, certain model features can be directly related to other observables from the data (e.g., item-total score regressions). A downside is that estimation can become tricky in case of incomplete

data (Eggen & Verhelst, 2011) and adaptive testing (Zwitser & Maris, 2015). So, in closing, our initial results are promising, but much more work is needed on model features (e.g., marginal functions for item responses and response times), estimation (e.g., standard errors) and model fit (e.g., generalized residuals).

References

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society, Series B (Methodological), 32*, 283–301.

Besag, J. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician), 24*(3), 179–195.

Bolsinova, M., Tijmstra, J., Molenaar, D., & De Boeck, P. (2017). Conditional independence between response time and accuracy: An overview of its possibles sources and directions for distinguishing between them. *Frontiers in Psychology, 8*. https://doi.org/10.3389/fpsyg.2017.00202

Eggen, T. J. H. M., & Verhelst, N. D. (2011). Item calibration in incomplete testing designs. *Psicológica, 32*(1), 107–132.

Haberman, S. J. (2007). The interaction model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 201–216). Springer.

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

Maris, G., & van der Maas, H. L. J. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika, 77*, 615-633. https://doi.org/10.1007/s11336-012-9288-y

Maris, G., & Bechger, T. (2021). Boltzmann machines as multidimensional item response theory models. Available via http://www.psyarxiv.com.

Maris, G., Bechger, T., Koops, J., & Partchev, I. (2022). dexter: Data management and analysis of tests [Computer software manual]. Available via https://dexter-psychometrics.github.io/dexter/.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum.

Müller, H. (1987). A Rasch model for continuous responses. *Psychometrika, 52*, 165–181. https://doi.org/10.1007/BF02294232

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Paedagogike Institut.

van der Linden, W. J., & Eggen, T. J. H. M. (1986). An empirical Bayesian approach to item banking. *Applied Psychological Measurement, 10*(4), 345–354.

van Rijn, P. W., & Ali, U. S. (2018a). A generalized speed-accuracy response model for dichotomous items. *Psychometrika, 83*, 109–131. https://doi.org/10.1007/s11336-017-9590-0

van Rijn, P. W., & Ali, U. S. (2018b). *SARM: A computer program for estimating speed-accuracy response models* (ETS Research Report RR-18-15). Educational Testing Service.

Verhelst, N. D. (2019). Exponential family models for continuous responses. In B. P. Veldkamp & C. Sluiter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 135–160). Springer.

Zwinderman, A. H. (1995). Pairwise parameter estimation in Rasch models. *Applied Psychological Measurement, 19*(4), 369–375.

Zwitser, R. J., & Maris, G. (2015). Conditional statistical inference with multistage testing designs. *Psychometrika, 80*(1), 65–84.

A Modified Method of Balancing Attribute Coverage in CD-CAT

Chia-Ling Hsu, Zi-Yan Huang, Chuan-Ju Lin, and Shu-Ying Chen

Abstract This study introduces a new attribute balancing method for cognitive diagnostic computerized adaptive testing (CD-CAT): the *modified attribute balancing index* (M-ABI). Based on simulation studies, using the M-ABI yielded acceptable measurement accuracy, ensured attribute coverage, and increased item bank utilization regardless of the item selection method, test length, or complexity of the Q-matrix structure. Overall, these results suggest the feasibility of using the M-ABI in CD-CAT to increase measurement precision, attribute coverage, and item usage, simultaneously.

Keywords Attribute coverage · Attribute balancing · Computerized adaptive testing · Cognitive diagnostic model

1 Introduction

Cognitive diagnostic computerized adaptive testing (CD-CAT) combines the psychometric properties of *cognitive diagnostic models* (CDMs) and CAT, the benefits of which drive increasing research in CD-CAT. Which identifies an examinee's mastery (versus non-mastery) of a set of fine-grained latent attributes (e.g., CDM, Rupp et al., 2010) by customizing each test to each examinee. As CD-CAT aims to accurately identify the attributes mastered by an examinee, its item selection algorithms must efficiently choose optimal items and/or balance the coverage of various attributes (*attribute balancing*; Cheng, 2010).

C.-L. Hsu (✉)
Hong Kong Examinations and Assessment Authority, Hong Kong SAR, China
e-mail: clhsu@hkeaa.edu.hk

Z.-Y. Huang · S.-Y. Chen
National Chung Cheng University, Minhsiung, Chiayi, Taiwan

C.-J. Lin
National University of Tainan, Tainan, Taiwan

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_12

Many past studies have shown how balancing attribute coverage for many items that measured *one* attribute affected the classification accuracy of examinees' mastery statues and ignored many items that measure *multiple* attributes. Extending the research, Sun et al. (2021) introduced an attribute balancing method for many items that measured multiple attributes. However, Sun et al.'s (2021) method does not guarantee balanced attribute coverage and ignores many items that measure a single attribute. These research results have shown that the published attribute balancing methods do improve the accuracy of examinees' mastery statues but do not uniformly utilize the item bank (e.g., they tend to select items that measured one or multiple attributes). Although accurately measuring an examinee's mastery status is the primary goal of CD-CAT, effective item use is also vital. Since building an item bank often entails a costly and time-consuming process of writing, reviewing, and pretesting the items, the existence of underused items in the item bank represents an undesirable resource waste. As a result, it is beneficial to have a cost-effective method of selecting items that can meet attribute coverage requirements, enhance classification accuracy, and boost item bank use at the same time.

To address this research gap, we introduce a new attribute balancing method: *the modified attribute balancing index* (M-ABI). Regardless of whether the item bank is composed of numerous items that measure a single or multiple attributes, the M-ABI increases both measurement precision and test efficiency. Specifically, the M-ABI (1) guarantees the balanced attribute coverage required for each attribute, (2) improves the measurement precision, (3) increases the selection of items that measure multiple attributes, and (4) thus increases the utilization of the item bank.

After briefly discussing CDM and item selection methods, we introduce the M-ABI. Then, we report on simulation studies comparing the M-ABI's performance to Cheng's (2010) and Sun et al.'s (2021) attribute balancing methods. Finally, we discuss the implications of this study for the use of M-ABI in CD-CAT.

2 The Deterministic Input Noisy Output "AND" Gate Model

This study uses the *deterministic input noisy output "AND" gate* (DINA) model (Junker & Sijtsma, 2001), in which an examinee must possess all *skills* (*latent attributes*) required by an item for a correct response. The probability of a correct response from examinee i on item j is as follows:

$$P\left(X_{ij}=1|\boldsymbol{\alpha}_i\right)=\left(1-s_j\right)^{\eta_{ij}}g_j{}^{1-\eta_{ij}}, \tag{1}$$

where the *mastery* (*latent class*) of examinee i is $\boldsymbol{\alpha}_i=(\alpha_{i1},\alpha_{i2},\ldots,\alpha_{iK})'$ and $\eta_{ij}=\prod_{k=1}^{K}\boldsymbol{\alpha}_{ik}^{q_{jk}}$ (q_{jk} indicates whether item j requires attribute k for a correct response), with the slipping parameter s_j and the guessing parameter g_j.

3 Item Selection Methods

Attribute balancing for item selection began with the *global discrimination index* (*GDI*, Cheng, 2010) using Kullback-Leibler (KL) information (Tatsouka, 2002; Tatsouka & Ferguson, 2003; Xu et al., 2003). We use KL information first, before using posterior-weighted KL (PWKL) information (Cheng, 2009) and then modified PWKL (MPWKL) information (Kaplan et al., 2015).

The KL information of item j for examinee i's provisional estimate of the latent class $\hat{\boldsymbol{\alpha}}$ is as follows:

$$KL_j\left(\hat{\boldsymbol{\alpha}}_i\right) = \sum_{c=1}^{2^K}\left[\sum_{x=0}^{1} P\left(X_j = x|\hat{\boldsymbol{\alpha}}_i\right)\log\left(\frac{P\left(X_j = x|\hat{\boldsymbol{\alpha}}_i\right)}{P\left(X_j = x|\boldsymbol{\alpha}_c\right)}\right)\right], \tag{2}$$

where $P\left(X_j = x|\hat{\boldsymbol{\alpha}}_i\right)$ and $P(X_j = x|\,\boldsymbol{\alpha}_c)$ are the probabilities of response x to item j given the interim estimate for examinee i's latent class and latent class c ($c = 1, 2, \ldots, 2^K$), respectively.

The PWKL information of item j for examinee i given the responses to t items is as follows:

$$PWKL_j\left(\hat{\boldsymbol{\alpha}}_i^{(t)}\right) = \sum_{c=1}^{2^K}\left\{\sum_{x=0}^{1}\left[P\left(X_j = x|\hat{\boldsymbol{\alpha}}_i^{(t)}\right)\log\left(\frac{P\left(X_j = x|\hat{\boldsymbol{\alpha}}_i^{(t)}\right)}{P\left(X_j = x|\boldsymbol{\alpha}_c\right)}\right)\right]\boldsymbol{\pi}^{(t)}\left(\boldsymbol{\alpha}_c\right)\right\}. \tag{3}$$

Given the responses to t items $\mathbf{x}^{(t)}$, KL information is weighted by the corresponding posterior probabilities of the latent classes (i.e., $\boldsymbol{\pi}^{(t)}(\boldsymbol{\alpha}_c) \propto \boldsymbol{\pi}(\boldsymbol{\alpha}_{c0})L(\mathbf{x}^{(t)}|\,\boldsymbol{\alpha}_c)$). The prior distribution for latent class c is $\boldsymbol{\pi}(\boldsymbol{\alpha}_{c0})$, the likelihood of $\mathbf{x}^{(t)}$ given latent class c is $L(\mathbf{x}^{(t)}|\,\boldsymbol{\alpha}_c)$, and the interim estimate for examinee i's latent class given t items is $\hat{\boldsymbol{\alpha}}_i^{(t)}$.

The MPWKL information for item j is as follows:

$$MPWKL_j\left(\hat{\boldsymbol{\alpha}}_i^{(t)}\right) = \sum_{d=1}^{2^K}\left\{\sum_{c=1}^{2^K}\sum_{x=0}^{1}\left[P\left(X_j = x|\boldsymbol{\alpha}_d\right)\log\left(\frac{P\left(X_j = x|\boldsymbol{\alpha}_d\right)}{P\left(X_j = x|\boldsymbol{\alpha}_c\right)}\right)\boldsymbol{\pi}^{(t)}\left(\boldsymbol{\alpha}_c\right)\right]\boldsymbol{\pi}^{(t)}\left(\boldsymbol{\alpha}_d\right)\right\}, \tag{4}$$

where the probability of response x to item j considering latent class d is $P(X_j = x|\,\boldsymbol{\alpha}_d)$ and the posterior probability of latent classes d is $\boldsymbol{\pi}^{(t)}(\boldsymbol{\alpha}_d)$.

4 A Modified Attribute Balancing Index

To better balance attribute coverage for an item bank with many multi-attribute items, Sun et al. (2021) proposed an attribute balancing method called *ratio of test length to the number of attributes* (*RTA*). The RTA for item j is as follows:

$$RTA_j = \frac{1}{1 + I\left(H \leq B_k\right) \sum_{v=1}^{V} I\left(\mathbf{q}_j = \mathbf{q}_v^*\right)}, H = \min\left(b_1, b_2, \ldots, b_K\right) \tag{5}$$

where B_k is the minimum number of items required to measure the kth attribute ($k = 1, 2, \ldots, K$), b_k is the number of previously selected items used to measure the kth attribute, $I(H \leq B_k)$ and $I\left(\mathbf{q}_j = \mathbf{q}_v^*\right)$ are the indicator functions, V is the number of previously selected items ($v = 1, 2, \ldots, V$), $\mathbf{q}_v^*$ is the q-vector of previously administered items, and $\mathbf{q}_j$ is the q-vector of unadministered items. As stated by Sun et al. (2021), the RTA is more likely to select items with attributes that differ from those of previously administered items, hence boosting the utilization of the item bank. However, the two elements, $I(H \leq B_k)$ and $\sum_{v=1}^{V} I\left(\mathbf{q}_j = \mathbf{q}_v^*\right)$, simultaneously determine the RTA; if one of them is 0, the other can be disregarded, which does not affect item selection. Thus, the RTA does not guarantee that each attribute fully meets the attribute coverage requirement.

To simultaneously increase the selection of multi-attribute items and fully meet the attribute coverage requirement, we modified the attribute balancing index (ABI, Cheng, 2010) to create the M-ABI, which releases the limit of the ABI from 0 when the kth attribute fulfils the attribute coverage requirement:

$$M - ABI_j = \sum_{k=1}^{K} \left(\frac{B_k - b_k}{B_k}\right) q_{jk}, \tag{6}$$

with all notations denoting the same meanings as in Eq. (5). As compared to $ABI_j = \prod_{k=1}^{K} (B_k - b_k / B_k)^{q_{jk}}$, the M-ABI has a slightly different form. First, q_{jk} is no longer an exponent but, rather, a multiplier of each $(B_k - b_k)/B_k$. Second, rather than calculating a *multiplicative* product of all $(B_k - b_k)/B_k$, we sum all the components. Larger differences among attributes and their corresponding unsatisfied attribute coverage requirements for each item yield both greater ABI and greater M-ABI values. However, when a b_k is equal to its corresponding B_k, ABI $= 0$, but M-ABI $\neq 0$; instead, M-ABI merely decreases. Thus, unlike ABI, M-ABI still allows these unused items that measure attribute k to be chosen and administered to the examinee.

5 Simulation Study

We used an item bank with 300 items that measured six attributes, generated the *g*- and *s*-parameters in the DINA model from Uniform (0.05, 0.25), and designed two Q-matrix structures (simple versus complex). The simple Q-matrix had 55% single-attribute items, and each item measured 20% the targeted attributes. By contrast, the complex Q-matrix (based on Sun et al., 2021) had 87% multi-attribute items. There were 30,000 examinees were generated, such that each examinee had a 50% chance of independently mastering each attribute.

Three item selection algorithms—KL, PWKL, and MPWKL—were employed with the ABI, RTA, and M-ABI methods. The fixed-length stopping rules stopped the CD-CAT for test lengths set at 10, 15, 20, and 25 items. To ensure satisfactory attribute balancing under various test lengths, the minimum number of items that measured each attribute was set to one, two, three, and four items for each of the four test lengths. Thus, 60%, 80%, 90%, and 96% of the test lengths, respectively, were considered as attribute balancing. The initial latent class was randomly generated from 64 latent classes for each examinee. The maximum likelihood estimation updated the examinees' provisional and final estimates of the latent classes. Thus, the simulation study had 72 conditions = 2 (Q-matrix structures) × 3 (item selection algorithms) × 3 (attribute balancing methods) × 4 (stopping rules). All conditions used identical simulated designs.

The proportion of examinees with correctly classified latent classes is called as the *classification accuracy rate* (CAR), and it indicates the accuracy of the attribute balancing method.

$$CAR = \frac{\sum_{i=1}^{N} I_{\hat{\boldsymbol{\alpha}}_i, \boldsymbol{\alpha}_i}}{N}, \tag{7}$$

with an indicator function $I_{\hat{\boldsymbol{\alpha}}_i, \boldsymbol{\alpha}_i}$ and the number of examinees N. If $\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i$, $I_{\hat{\boldsymbol{\alpha}}_i, \boldsymbol{\alpha}_i} = 1$; otherwise, $I_{\hat{\boldsymbol{\alpha}}_i, \boldsymbol{\alpha}_i} = 0$. The attribute coverage of the proportion of examinees whose tests satisfied the attribute balancing criterion indicates the efficacy of the attribute balancing method. Meanwhile, the number of items used across all examinees (so-called *item usage*) indicates the efficiency of the attribute balancing method. It would be expected that, as compared with the ABI and RAT methods, the proposed M-ABI will meet the three criteria simultaneously, enabling it cost-effective.

6 Results

Table 1 shows that, regardless of the item selection method, Q-matrix structure, or test length, both M-ABI and ABI ensured that all the tests satisfied the attribute

Table 1 Percentage of examinees' tests meet attribute coverage requirement

Item selection	Q-matrix	Attribute balancing	Test length			
			10	15	20	25
KL	Simple	ABI	1.00	1.00	1.00	1.00
		RTA	0.97	0.78	0.63	0.63
		M-ABI	1.00	1.00	1.00	1.00
	Complex	ABI	1.00	1.00	1.00	1.00
		RTA	0.99	0.80	0.68	0.66
		M-ABI	1.00	1.00	1.00	1.00
PWKL	Simple	ABI	1.00	1.00	1.00	1.00
		RTA	1.00	0.94	0.88	0.82
		M-ABI	1.00	1.00	1.00	1.00
	Complex	ABI	1.00	1.00	1.00	1.00
		RTA	1.00	0.95	0.87	0.78
		M-ABI	1.00	1.00	1.00	1.00
MPWKL	Simple	ABI	1.00	1.00	1.00	1.00
		RTA	1.00	0.96	0.84	0.73
		M-ABI	1.00	1.00	1.00	1.00
	Complex	ABI	1.00	1.00	1.00	1.00
		RTA	1.00	0.94	0.78	0.66
		M-ABI	1.00	1.00	1.00	1.00

Note. *KL* Kullback–Leibler information, *PWKL* posterior-weighted KL, *MPWKL* modified PWKL, *NAB* no attribute balancing, *ABI* attribute balancing index, *RTA* ratio of test length to the number of attributes, *M-ABI* modified ABI

coverage criterion but RAT did not. For strict attribute balancing constraints (e.g., 90% or 96% of the test lengths were considered as attribute balancing), the M-ABI, ABI, and RTA methods showed similar results across all simulation conditions (their differences in CARs were < .05; see Table 2). For lenient attribute balancing constraints (e.g., 60% of the test lengths were considered as attribute balancing), the ABI often performed slightly better than both RTA (by 0–.05) and M-ABI (by .04–.09, see Table 2). Figure 1 shows that the M-ABI had higher efficiency than both ABI and RTA, using 16.7% more items than the ABI and 7.3% more items than the RTA on average across conditions.

7 Discussion

This study aims to propose a cost-efficient method for considering classification accuracy, attribute coverage efficacy, and item bank utilization simultaneously. Thus, we introduced a new method, *the modified attribute balancing index* (M-ABI), and compared it to two other methods (ABI and RTA) regarding these three criteria. Overall, the M-ABI satisfied the attribute balancing requirement (like the ABI but

Table 2 Classification accuracy rate of latent classes

Item selection	Q-matrix	Attribute balancing	Test length			
			10	15	20	25
KL	Simple	ABI	0.60	0.76	0.86	0.92
		RTA	0.59	0.77	0.86	0.91
		M-ABI	0.53	0.73	0.84	0.90
	Complex	ABI	0.58	0.75	0.83	0.87
		RTA	0.54	0.73	0.82	0.86
		M-ABI	0.49	0.69	0.80	0.86
PWKL	Simple	ABI	0.74	0.92	0.96	0.98
		RTA	0.69	0.89	0.96	0.99
		M-ABI	0.67	0.88	0.96	0.98
	Complex	ABI	0.70	0.88	0.93	0.96
		RTA	0.65	0.85	0.94	0.97
		M-ABI	0.63	0.83	0.92	0.96
MPWKL	Simple	ABI	0.80	0.93	0.96	0.97
		RTA	0.80	0.93	0.98	0.99
		M-ABI	0.76	0.91	0.97	0.99
	Complex	ABI	0.77	0.88	0.92	0.95
		RTA	0.77	0.90	0.96	0.98
		M-ABI	0.73	0.89	0.95	0.97

Note. *KL* Kullback–Leibler information, *PWKL* posterior-weighted KL, *MPWKL* modified PWKL, *NAB* no attribute balancing, *ABI* attribute balancing index, *RTA* ratio of test length to the number of attributes, *M-ABI* modified ABI

not the RTA), had similar CAR given strict attribute balancing constraints, had slightly less CAR given lenient attribute balancing constraints (ABI $\geq$ RTA > M-ABI), and was much more efficient in terms of item utilization. The M-ABI has a slightly lower CAR than ABI/RTA because it uses more items for administration, which may not contribute well to the objective function optimized in the CD-CAT algorithms. Nonetheless, as compared with ABI and RTA, M-ABI had acceptable classification accuracy, ensured balanced attribute coverage, and increased item usage; thus, when performed simultaneously with CD-CAT, it was the most cost-effective method when one considers accuracy, attribute coverage efficacy, and efficiency.

This study's limitations include its few manipulations and absence of real-world data. First, we manipulated only a few factors/levels in the simulation studies, so future studies can manipulate more such factors or levels, such as by adopting different CDMs, different Q-matrix structures, more attributes, other advanced item selection algorithms (e.g., Wang, 2013; Zheng & Chang, 2016), or different test termination rules (e.g., Hsu et al., 2013; Tatsouka, 2002). Furthermore, M-ABI performance might differ under different practical constraints, such as test security and content balancing control, so future studies can examine such constraints.

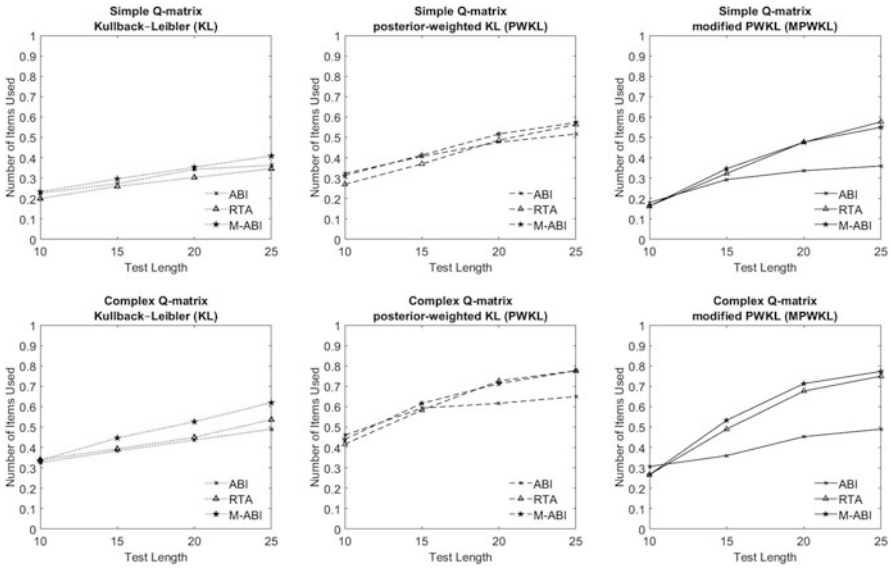


Fig. 1 Item bank utilization. (*Note*. *ABI* attribute balancing index, *RTA* ratio of test length to the number of attributes, *M-ABI* modified ABI)

References

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*, 619–632. https://doi.org/10.1007/s11336-009-9123-2

Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: The modified maximum global discrimination index method. *Educational and Psychological Measurement, 70*, 902–913. https://doi.org/10.1177/0013164410366693

Hsu, C.-L., Wang, W.-C., & Chen, S.-Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement, 37*, 563–582. https://doi.org/10.1177/0146621613488642

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. https://doi.org/10.1177/01466210122032064

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for computerized adaptive testing. *Applied Psychological Measurement, 39*, 167–188. https://doi.org/10.1177/0146621614554650

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications* (The statistical structure of core DCMs). Guilford.

Sun, X., Andersson, B., & Xin, T. (2021). A new method to balance measurement accuracy and attribute coverage in cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement, 45*, 463–476. https://doi.org/10.1177/01466216211040489

Tatsouka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of Royal Statistical Society: Series C (Apply Statistics), 51*, 337–350. https://doi.org/10.1111/1467-9876.00272

Tatsouka, C., & Ferguson, T. (2003). Sequential classification on partially ordered sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65*, 143–157. http://www.jstor.org/stable/3088831

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement, 73*, 1017–1035. https://doi.org/10.1177/0013164413498256

Xu, X., Chang, H., & Douglas, J. (2003). *A simulation study to compare CAT strategies for cognitive diagnosis*. Paper presented at the Annual Meeting of the American Education Research Association, Chicago, IL.

Zheng, C., & Chang, H. H. (2016). High-efficiency response distribution-based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement, 40*, 608–624. https://doi.org/10.1177/0146621616665196

Resolving the Test Fairness Paradox by Reconciling Predictive and Measurement Invariance

Safir Yousfi

Abstract Until recently, the dominant approach to establish that a psychological or educational test is fair with respect to a demographic characteristic like gender or age was to show that predictive invariance holds (e.g. identical regression of the criterion on the test scores). In last decade, the claim of some psychometricians that measurement invariance should be regarded as a major prerequisite for test fairness had some impact. Both criteria for test fairness are now required by common standards for educational and psychological testing.

However, it has been shown that predictive invariance and measurement invariance are incompatible concepts and cannot hold simultaneously in realistic settings. While psychometricians concluded that an explicit choice between these approaches has to be made, test developers and test users seem to neglect the incompatibility and follow one of the approaches without explanation.

A psychometric approach to test fairness is suggested that resolves the incompatibility of predictive and measurement invariance by adopting the key ideas behind both competing concepts of test fairness. Within this framework, latent predictive invariance and measurement invariance are both considered as basic requirements for test fairness in realistic settings. Additional requirements lead to more stringent concepts of fairness that are necessary for fairness considerations in case of multidimensionality.

Keywords Test fairness · Measurement invariance · Predictive invariance · Latent variables

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

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_13

1 Introduction

It has been proven that measurement invariance and predictive invariance are incompatible requirements under almost all circumstances (Meredith & Millsap, 1992; Millsap & Meredith, 1992; Millsap, 1997, 2007, 2011). Hence, it is an unattainable goal to meet both most common statistical requirements for test fairness. Consequently, Borsboom et al. (2008) concluded that an explicit decision between both concepts is inevitable and recommended to prefer measurement invariance. In the remainder, a different approach for escaping from this dilemma will be outlined.

2 Recap of Predictive and Measurement Invariance and Their Relations

Definition (Predictive Invariance) Predictive invariance holds if for each value of the variable $\boldsymbol{X}$, the conditional distribution of the criterion variable C does not depend on the value of group variable $\boldsymbol{G}$ (cf. Cleary, 1968; Millsap, 2007):

$$(\boldsymbol{G} \perp\!\!\!\perp C) \mid \boldsymbol{X}$$

Definition (Measurement Invariance) Measurement invariance[1] holds if for each value of the vector of latent variables $\boldsymbol{\theta_X}$ the conditional distribution of the variable $\boldsymbol{X}$ does not depend on the value of group variable $\boldsymbol{G}$ (cf. Meredith & Millsap, 1992; Millsap, 2007):

$$(\boldsymbol{G} \perp\!\!\!\perp \boldsymbol{X}) \mid \boldsymbol{\theta_X}$$

$\boldsymbol{X}$ can be test response pattern and/or a function thereof.

Meredith and Millsap (1992) have shown that $(\boldsymbol{G} \perp\!\!\!\perp \boldsymbol{X}) \mid \boldsymbol{\theta_X}$ (measurement invariance) implies $(\boldsymbol{G} \perp\!\!\!\perp C) \mid \boldsymbol{X}$ (predictive invariance) in case of Bayesian sufficiency, i.e. if

$$(C \perp\!\!\!\perp \boldsymbol{\theta_X}) \mid \boldsymbol{X}$$

Meredith and Millsap (1992) emphasize that $(C \perp\!\!\!\perp \boldsymbol{\theta_X}) \mid \boldsymbol{X}$ (Bayesian sufficiency) would hardly be a met in practical applications as it requires that measurement errors of $\boldsymbol{X}$ (as a measure of $\boldsymbol{\theta_X}$) would not attenuate but contribute unrestrictedly to the association of $\boldsymbol{\theta_X}$ and C.

[1] The respective constellation is sometimes referred to as strict or absolute measurement invariance which contrast to other weaker forms of measurement invariance. In this paper we do not use any variations of measurement invariance but refer always to this concept.

In contrast, if C and $\boldsymbol{X}$ are locally independent in each group, i.e. $(C \perp\!\!\!\perp \boldsymbol{X}) \mid (\boldsymbol{\theta_X}, \boldsymbol{G})$, and there are group differences on the latent variable (i.e., $\boldsymbol{G} \not\perp\!\!\!\perp \boldsymbol{\theta_X}$) then measurement invariance (i.e., $(\boldsymbol{G} \perp\!\!\!\perp \boldsymbol{X}) \mid \boldsymbol{\theta_X}$) and predictive invariance (i.e., $(\boldsymbol{G} \perp\!\!\!\perp C) \mid \boldsymbol{X}$) are mutually exclusive (unless some trivial regularity assumptions are violated; Meredith & Millsap, 1992). Hence, it is an unattainable goal to meet both most common statistical requirements for test fairness. Borsboom et al. (2008) concluded that an explicit decision between both concepts is inevitable and recommended to prefer measurement invariance. In the following chapter, a different approach for escaping from this dilemma will be outlined.

3 Fairness Concepts Based on Latent Variables

The key idea behind predictive invariance is that disparate impact (substantial group differences in the consequences of a procedure, e.g. selection rates) can be justified (only) if it can be attributed to a trait that is shown to be predictive for a respective outcome (e.g. success on the job). This line of reasoning relies on the (implicit) assumption that disparate treatment of persons that differ (only) in the respective trait is justified. For example, lower admission rates for black applicant need not necessarily considered as unfair, if they can be attributed to lower abilities (instead of direct discrimination based on their race). These lower abilities result in lower aptitude and lower aptitude is generally regarded as a justified reason for disparate treatment.

However, observed test scores are at best (statistically) unbiased estimators for the respective latent variable. The predictive and explanatory power of the observed score stems from the underlying latent variable and is attenuated by measurement error. As a consequence, the posterior distributions (on the latent variable) of persons with the same observed score on a trait measure generally differs between groups if there are group differences on the trait. This leads to violations of predictive invariance even if the latent trait is regarded as criterion variable. Such violations of predictive invariance should not be considered as a violation of principles of fairness as it does not result in disparate treatment of members of different groups with the same latent trait score. Principles of fairness would only be violated if group differences on the criterion variable are expected for persons with the same latent trait value as these differences would undercut the premise that disparate impact (group differences in the consequences of a procedure) could be justified by differences on the respective trait.

In conclusion, predictive invariance is generally not suited to justify disparate impact or disparate treatment of members of different groups based on their observed trait measures. Requiring predictive invariance for the latent variable as predictor of the respective criterion seems to be an obvious alternative fairness requirement. Moreover, the incompatibility of predictive invariance and measurement invariance applies only the observed test scores as predictor of the criterion.

There is no incompatibility of latent predictive invariance with measurement invariance. As any violation of latent predictive invariance and measurement invariance inevitably leads to fairness issues it seems straightforward to integrate both concepts in a comprehensive psychometric concept of fairness.

Definition (Weak Fairness) Using a function f of the test response $\boldsymbol{X}$ as indicator/predictor of C is weakly fair with regard to characteristic $\boldsymbol{G}$ at (all levels of) $\boldsymbol{\theta_X}$, if

a. $(C \perp\!\!\!\perp f(\boldsymbol{X})) \mid (\boldsymbol{\theta_X}, \boldsymbol{G})$ (group-wise local independence of $f(\boldsymbol{X})$ and C)
b. $(\boldsymbol{G} \perp\!\!\!\perp f(\boldsymbol{X})) \mid \boldsymbol{\theta_X}$ (measurement invariance of $f(\boldsymbol{X})$)
c. $(\boldsymbol{G} \perp\!\!\!\perp C) \mid \boldsymbol{\theta_X}$ (predictive invariance of $\boldsymbol{\theta_X}$)

Please note that weak fairness boils down to measurement invariance of $f(\boldsymbol{X})$ in case of $C = \boldsymbol{\theta_X}$ (operational definition of the intended target of measurement). In this case, $\boldsymbol{X}$ might be the test response pattern and $f(\boldsymbol{X})$ might an estimator of $\boldsymbol{\theta_X}$, e.g. the maximum likehood, weighted likelihood, or a Bayesian estimator (without group specific prior) or simply the sum score.

Please note also, that weak fairness boils down to predictive invariance of $f(\boldsymbol{X})$, if $f(\boldsymbol{X}) = \boldsymbol{\theta_X}$, i.e. if $f(\boldsymbol{X})$ is a perfectly reliable unbiased estimator of $\boldsymbol{\theta_X}$.

For theoretical purposes C might be a platonic true score, i.e. a latent variable that is defined conceptually and not necessarily by a measurement model of observed variables.

Weak fairness refers always to a specific grouping of the population by the group variable $\boldsymbol{G}$. Moreover, weak fairness refers always to a specific scoring f of the test response pattern $\boldsymbol{X}$. Consequently, item DIF (i.e. violations of measurement invariance for elements of $\boldsymbol{X}$) does not necessarily lead to violations of measurement invariance of $f(\boldsymbol{X})$, e.g. if DIF cancels out across the items.

Weak fairness ensures that (for each level of $\boldsymbol{\theta_X}$) statistical inference from $\boldsymbol{X}$ on C by means of f does not depend on $\boldsymbol{G}$. Hence, it ensures fairness on each level of $\boldsymbol{\theta_X}$. This is a trivial consequence of the local independence assumption which requires that $f(\boldsymbol{X})$ is not informative with respect to C on all levels of $\boldsymbol{\theta_X}$ (but may be informative across levels of $\boldsymbol{\theta_X}$).

However, if the (expected) error of inference from $f(\boldsymbol{X})$ on C (i.e. $C - f(\boldsymbol{X})$), depends on $\boldsymbol{\theta_X}$ then group differences on C might not be adequately reflected by $f(\boldsymbol{X})$. In extreme cases, group differences on $f(\boldsymbol{X})$ might even have different sign/direction than group differences on C. Moreover, a scoring f is always weakly fair if $f(\boldsymbol{X})$ is an arbitrary constant that takes the same value regardless of $\boldsymbol{X}$. Weak fairness would also hold, if only those elements of a $\boldsymbol{X}$ that are completely unrelated to $\boldsymbol{\theta_X}$ (e.g. 2-pl items with zero discrimination) are considered by $f(\boldsymbol{X})$ while elements of $\boldsymbol{X}$ that are related to $\boldsymbol{\theta_X}$ are ignored. To avoid such undesired properties, it seems necessary to tighten the concept of weak fairness by additional fairness requirements, if $f(\boldsymbol{X})$ and C are on an ordinal or interval scale. Then it is expedient to require that $f(\boldsymbol{X})$ reflects differences of C due to differences in $\boldsymbol{\theta_X}$ in order to keep the claim tenable that group difference in $f(\boldsymbol{X})$ can be justified by

group difference in $\boldsymbol{\theta_X}$ that are predictive for C. Otherwise the conceptual basis that underlies (latent) predictive invariance would be undercut.

Definition (Substantial Fairness) A weakly fair test use is called substantially fair, if the following condition holds for each pair of values ***a,b*** of the latent variable

$$E\,(C|\boldsymbol{\theta_X} = \boldsymbol{a}) \leq E\,(C|\boldsymbol{\theta_X} = \boldsymbol{b}) \implies E\,(f\,(\boldsymbol{X})\,|\boldsymbol{\theta_X} = \boldsymbol{a}) \leq E\,(f\,(\boldsymbol{X})\,|\boldsymbol{\theta_X} = \boldsymbol{b})$$

Substantial fairness ensures that differences in $\boldsymbol{\theta_X}$ that are consequential for the expected value of C are also reflected in expected value of $f(\boldsymbol{X})$. However, substantial fairness is not invariant to strictly monotone transformations of $f(\boldsymbol{X})$ and focuses only on expectations and neglects other properties of the distributions of C and $f(\boldsymbol{X})$. Hence, it seems necessary to tighten the concept of fairness further.

Definition (Essential Fairness) A weakly fair test use is called essentially fair, if the following condition holds for each pair ***a,b*** of values the latent variable

$$(C|\boldsymbol{\theta_X} = \boldsymbol{a}) \preccurlyeq (C|\boldsymbol{\theta_X} = \boldsymbol{b}) \quad \implies \quad (f\,(\boldsymbol{X})\,|\boldsymbol{\theta_X} = \boldsymbol{a}) \preccurlyeq (f\,(\boldsymbol{X})\,|\boldsymbol{\theta_X} = \boldsymbol{b})$$

$(A|\,B) \preccurlyeq (D|\,E)$ denotes "less in the usual stochastic order", i.e. $P(A > x|\,B) \leq P(D > x\,|\,E)$ for each value of x.

Essential fairness ensures that the usual stochastic orderings of the conditional distributions of C and $f(\boldsymbol{X})$ are consistent (on the whole domain of $\boldsymbol{\theta_X}$). In other in words: If a value $\boldsymbol{b}$ of the latent trait $\boldsymbol{\theta_X}$ is associated with higher values of the criterion C than another value $\boldsymbol{a}$, then $\boldsymbol{b}$ will also be associated with higher values of $f(\boldsymbol{X})$, i.e. $f(\boldsymbol{X})$ reflects differences of C due to differences in $\boldsymbol{\theta_X}$.

However, essential fairness does not preclude underprediction and overprediction of groups that differ with respect to their distribution on $\boldsymbol{\theta_X}$. If overprediction happens on lower levels of $\boldsymbol{\theta_X}$, for example, then the magnitude of the (mean) difference between a low scoring group to another group with average or above average (mean) values of $\boldsymbol{\theta_X}$ would not be adequately reflected in the values of $f(\boldsymbol{X})$. To prevent bias on the group level the concept of weak fairness can be tightened further by referring to difference of C and $f(\boldsymbol{X})$. This difference $\boldsymbol{\varepsilon} \coloneqq \boldsymbol{crit} - f(\boldsymbol{X})$ might be called error of prediction, if C is an observed variable or error of interpretation if C is a hypothetical variable that is defined conceptually (plantonic true score).

Definition (Strong Fairness) Weakly fair test use is called strongly fair if $E(\boldsymbol{\varepsilon}|\,\boldsymbol{\theta_X}) = E(\boldsymbol{\varepsilon})$ (conditional regressive independence of $\boldsymbol{\varepsilon}$ and $\boldsymbol{\theta_X}$)

Definition (Strict Fairness) Weakly fair test use is called strictly fair if $\boldsymbol{\varepsilon} \perp\!\!\!\perp \boldsymbol{\theta_X}$ (stochastic independence of $\boldsymbol{\varepsilon}$ and $\boldsymbol{\theta_X}$).

Definition (Absolute Fairness) Absolutely fair test use is strictly fair test use with $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$

Please note that the additional requirements for substantial, essential, strong, strict and absolute fairness do not refer to $\boldsymbol{G}$. Consequently, neither of these requirements are directly relevant for fairness with respect to $\boldsymbol{G}$. As fairness on each level of $\boldsymbol{\theta_X}$ is already guaranteed by weak fairness, these additional requirements aim at fairness with respect to $\boldsymbol{\theta_X}$ to prevent indirect bias as consequence of group differences on $\boldsymbol{\theta_X}$. They ensure that $f(\boldsymbol{X})$ reflects differences of C due to differences in $\boldsymbol{\theta_X}$. The required level of fidelity in reflecting these differences increases from substantial to essential to strong to strict and finally to absolute fairness.

Only absolute fairness prevents against any over- and underestimation for all levels of $\boldsymbol{\theta_X}$. Strict fairness ensures absolutely homogenous over- or underestimation if $E(\boldsymbol{\varepsilon}) \neq \boldsymbol{0}$ and prevents against any bias with respect to $\boldsymbol{G}$. Strong fairness requires only that the expected over- or underestimation is the same for all values of $\boldsymbol{\theta_X}$. However, the precision of the prediction of C might depend on $\boldsymbol{\theta_X}$ which might lead to minor violations of fairness on the group level in case of group differences on $\boldsymbol{\theta_X}$. If C or $f(\boldsymbol{X})$ are only on an ordinal scale, then $\boldsymbol{\varepsilon}$ is generally not meaningful which undercuts the conceptual basis of strong, strict and absolute fairness. In contrast, essential fairness is meaningful if C and/or $f(\boldsymbol{X})$ are on an ordinal scale. Consequently, essential fairness is robust against monontone transformations of C and $f(\boldsymbol{X})$. If essential fairness is violated then substantial fairness cannot be expected to be robust against all monotone transformations of C and $f(\boldsymbol{X})$. If substantial fairness is violated for a weakly fair test, then local fairness still holds on each level of $\boldsymbol{\theta_X}$. However, $f(\boldsymbol{X})$ does not imply the same order on $\boldsymbol{\theta_X}$ as C does. Consequently, the order of different groups (values of $\boldsymbol{G}$) with respect to $f(\boldsymbol{X})$ might be inconsistent with the order with respect to C which must be considered as a severe violation of fairness on the group level.

4 Fairness Concepts Based on Manifest Variables

If weak fairness is violated because of violations of the local independence assumption (i.e. measurement invariance and latent predictive variance still hold) then the claim that the explanatory power of the test scores stems only from respective latent trait but not from measurement error of $\boldsymbol{\theta_X}$ is not tenable as these measurement errors seem to be related to C. Such constellations are likely to occur if measurement errors of $\boldsymbol{\theta_X}$ do not only reflect transient variation of behavior but enduring changes as a result of the testing experience (e.g. learning).

In order to illustrate the effects of violations of local independency, we will refer to a hypothetical dating platform that has one matching algorithm $g(\boldsymbol{Y})$ that relies only on information $\boldsymbol{Y}$ that is available before potential mating partners interact and another algorithm $f(\boldsymbol{X})$ that relies on information $\boldsymbol{X}$ that is gathered after the first interaction of potential partners. Suppose that the same latent trait variable $\boldsymbol{\theta_X} = \boldsymbol{\theta_Y}$ underlies the observed scores $f(\boldsymbol{X})$ and $g(\boldsymbol{Y})$. $\boldsymbol{\theta_X}$ might be the (a-priori) mating

probability of potential partners. Measurement errors of $\boldsymbol{Y}$ and $g(\boldsymbol{Y})$ are most likely transient whereas random variation of $\boldsymbol{X}$ and $f(\boldsymbol{X})$ might reflect permanent effects of affection which might result in mating C. If the algorithm $f(\boldsymbol{X})$ is a (nearly) perfect measure of the state of potential couples after their first interaction, then it seems at least plausible that knowledge of $\boldsymbol{\theta_X}$ would not have any incremental value over $f(\boldsymbol{X})$ for the prediction of mating, i.e. Bayesian sufficiency of $f(\boldsymbol{X})$ with respect to $\boldsymbol{\theta_X}$ for the prediction of C would hold. If we suppose further, that measurement invariance of $f(\boldsymbol{X})$ holds for straight vs. queer potential couples ($\boldsymbol{G}$), then we could infer that predictive invariance of $f(\boldsymbol{X})$ holds (Meredith & Millsap, 1992; Millsap, 2007). However, the irrelevance of $\boldsymbol{\theta_X}$ as incremental predictor of C follows directly from Bayesian sufficiency and does not rely on measurement variance. As measurement errors of $f(\boldsymbol{X})$ as a measure of $\boldsymbol{\theta_X}$ (a-priori mating probability) do not attenuate but contribute to the prediction of C there is no need to require measurement invariance as prerequisite for manifest fairness (of post-hoc mating predictions after the first interaction). Therefore, in the remainder only Bayesian sufficiency but not measurement invariance would be required for manifest fairness on levels of the observed $f(\boldsymbol{X})$.

Definition (Weak Manifest Fairness) Weak manifest fairness (of using a function f of the test response $\boldsymbol{X}$ as indicator/predictor of C) holds (at all levels of $f(\boldsymbol{X})$), if the following conditions are met:

1. $(C \perp\!\!\!\perp \boldsymbol{\theta_X}) \mid f(\boldsymbol{X})$ (Bayesian sufficiency of $f(\boldsymbol{X})$)
2. $(\boldsymbol{G} \perp\!\!\!\perp C) \mid f(\boldsymbol{X})$ (predictive invariance of $f(\boldsymbol{X})$).

Definition (Substantial Manifest Fairness) Substantial manifest fairness is given, if weak manifest fairness holds and if the following condition holds for each pair $\boldsymbol{a,b}$ of values of $f(\boldsymbol{X})$

$$\boldsymbol{E}\,(\boldsymbol{C} \mid f\,(\boldsymbol{X}) = \boldsymbol{a}) \leq \boldsymbol{E}\,(\boldsymbol{C} \mid f\,(\boldsymbol{X}) = \boldsymbol{b}) \quad \Longrightarrow \quad \boldsymbol{a} \leq \boldsymbol{b}$$

Definition (Essential Manifest Fairness) A weakly manifestly fair test use is called essentially manifestly fair, if the following condition holds for each pair of values $\boldsymbol{a,b}$ of $f(\boldsymbol{X})$:

$$(C \mid f\,(\boldsymbol{X}) = \boldsymbol{a}) \preccurlyeq (C \mid f\,(\boldsymbol{X}) = \boldsymbol{b}) \quad \Longrightarrow \quad \boldsymbol{a} \leq \boldsymbol{b}$$

Definition (Strong Manifest Fairness) Weak manifest fairness is called strong manifest fairness if $E(\boldsymbol{\varepsilon} \mid f(\boldsymbol{X})) = E(\boldsymbol{\varepsilon})$ (conditional regressive independence of $\boldsymbol{\varepsilon}$ given $f(\boldsymbol{X})$)

Definition (Strict Manifest Fairness) Weak manifest fairness is called strict manifest fairness if $\boldsymbol{\varepsilon} \perp\!\!\!\perp f(\boldsymbol{X})$ (stochastic independence of $\boldsymbol{\varepsilon}$ and $f(\boldsymbol{X})$).

Definition (Absolute Manifest Fairness) Absolute manifest fairness is strict manifest fairness with $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$.

5 Comparison of Fairness Concepts for Latent and Manifest Variables

The concepts for manifest fairness mimic the concepts of latent fairness. The key difference is that latent fairness concepts require fairness on levels of $\boldsymbol{\theta}_{\boldsymbol{X}}$, while manifest fairness concepts require fairness on levels of $f(\boldsymbol{X})$. If $f(\boldsymbol{X})$ must be considered as fallible measure of a trait that is to be measured, then latent concepts of fairness should be applied to guarantee fairness for the units of measurement. If the random variation of $f(\boldsymbol{X})$ is also considered as a feature that is to be measured, then manifest fairness concepts should be applied for fairness evaluations. In most applications latent fairness concepts are much better compatible with the purpose of measurement. In contrast, it is hard to imagine settings where manifest fairness concepts fit to the purpose of measurement (cf. Meredith & Millsap, 1992; Millsap, 2007).

It is interesting to apply the fairness concepts to settings where $\mathbf{G}$ is constant and does not vary across the units of measurement. Then weak fairness boils down to local independence of $\mathbf{X}$ and C and manifest weak fairness boils down to Bayesian sufficiency. In this case, the other latent versions of fairness would describe the degree to which $f(\boldsymbol{X})$ reflects differences of C due to differences in $\boldsymbol{\theta}_{\boldsymbol{X}}$, while the other manifest version of fairness would describe the degree to which $f(\boldsymbol{X})$ reflects differences in C.

6 Discussion

Borsboom (2006) and Millsap (2007) referred to the apparent neglect of the work of Millsap (1997) and the focus on analyzing predictive invariance (instead of measurement invariance) for evaluating potential test bias as a paradigmatic example of a lack of impact of psychometric insights on psychological research and practice. Ten years later, Putnick and Bornstein (2016) reported an exponential growth in the number of studies that analyze the measurement invariance, i.e. the approach recommended by Borsboom et al. (2008). More recently, Han et al. (2019, p. 1484) concluded that "relative to predictive invariance, MI [measurement invariance] has been investigated more frequently and more rigorously in recent years . . . ", with some exception like " . . . assessment research related to personnel se[le]ction". However, even in test bias research that relies on analyses of predictive invariance, there is an increasing awareness that these analyses can be distorted by statistical artifacts due to measurement error (and in case of analyses based on correlational analyses also by range restriction). Attenuation corrections are recommended to adjust for these distortions (cf. Aguinis et al., 2010; Berry & Zhao, 2015). Interestingly, these corrections implicitly accomplish that the true score of

a test is considered as predictor instead of the observed score which is analogous to requiring latent predictive invariance instead of observed predictive invariance as prerequisite for fair test use. So far it seems to be unnoticed that attenuation corrections remove the incompatibility of predictive invariance and measurement that was described in the psychometric literature, i.e. both major statistical fairness approaches become compatible even in settings with local stochastic independence. This should not only be considered as an opportunity but rather as an obligation to apply both major psychometric approaches in future research and practice if fairness is to be evaluated. The fact that an integrated evaluation of fairness has hardly been tried in the literature (see Han et al., 2019) might at least in part be attributable to the discouraging message from the psychometric literature that such an endeavor would inevitably (or at least in realistic settings) lead to a violation of test fairness with respect to at least one of both major fairness criteria (Millsap, 1997, 2007; Borsboom et al., 2008).

One implication of the present work is that relying exclusively on analyses of measurement invariance is only adequate if the latent variable itself is considered as the ultimate criterion. This might be the case when the test has perfect content validity. However, this is rarely the case whenever inferences on or predictions of behavior in real life settings (that do not match exactly the test setting) are the purpose of measurement. On the other hand, relying only on latent predictive invariance without considering measurement invariance is insufficient for establishing test fairness as measurement bias leads inevitably to disparate treatment for test takers of different groups with the same value of the latent variable.

Another key message of the current work is that measurement invariance and (latent) predictive invariance jointly establish only a weak form of fairness that does not even preclude that higher expected test scores are associated with lower expected scores on the criterion. The additional requirements for substantial fairness might appear trivial in case of a one-dimensional latent variable. In case of multidimensionality substantial fairness (and the other concepts beyond weak fairness) requires that the relative importance of the dimensions of the latent variable for the criterion C are perfectly reflected in the test score (i.e. proportional regression weights for the dependent variable C and the dependent variable $f(\boldsymbol{X})$ if the associations are linear and additive). This line of reasoning indicates that the introduced concepts beyond weak fairness offer a promising framework to analyze fairness issues that result from omitted variables (cf. Sackett et al., 2003). In general, fairness with regard to the full vector of latent variables does not imply fairness with regard to a reduced or shortened vector of latent variables. Addressing these issues is important for many real-life applications of test scores where relevant criteria are often multidimensional. A full analysis of this topic is beyond the scope of this chapter.

References

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology, 95*, 648–680. https://doi.org/10.1037/a0018714

Berry, C. M., & Zhao, P. (2015). Addressing criticisms of existing predictive bias research: Cognitive ability test scores still overpredict African Americans' job performance. *Journal of Applied Psychology, 100*, 162–179.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*, 425–440. https://doi.org/10.1007/s11336-006-1447-6

Borsboom, D., Romeijn, J. W., & Wicherts, J. M. (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods, 13*, 75–98. https://doi.org/10.1037/1082-989X.13.2.75

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5*, 115–124. https://doi.org/10.1111/j.1745-3984.1968.tb00613.x

Han, K., Colarelli, S. M., & Weed, N. C. (2019). Methodological and statistical advances in the consideration of cultural diversity in assessment: A critical review of group classification and measurement invariance testing. *Psychological Assessment, 31*, 1481–1496. https://doi.org/10.1037/pas0000731

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57*, 289–311.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2*, 248–260. https://doi.org/10.1037/1082-989X.2.3.248

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika, 72*, 461–473. https://doi.org/10.1007/s11336-007-9039-7

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Millsap, R. E., & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias. *Applied Psychological Measurement, 16*, 389–402.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Sackett, P. R., Laczo, R. M., & Lippe, Z. P. (2003). Differential prediction and the use of multiple predictors: The omitted variables problem. *Journal of Applied Psychology, 88*, 1046–1056.

The Plausibility and Feasibility of Remedies for Evaluating Structural Fit

Graham G. Rifenbark and Terrence D. Jorgensen

Abstract Various structural fit indices (SFIs) have been proposed to evaluate the structural component of a structural equation model (SEM). Decomposed SFIs treat estimated latent (co)variances from an unrestricted confirmatory factor analysis (CFA) as input data for a path model, from which standard global fit indices are calculated. Conflated SFIs fit a SEM with both measurement and structural components, comparing its fit to orthogonal and unrestricted CFAs. Sensitivity of conflated SFIs to the same structural misspecification depends on standardized factor loadings, but decomposed SFIs have inflated Type-I error rates when compared to rule-of-thumb cutoffs, due to treating estimates as data. We explored whether two alternative approaches avoid either shortcoming by separating the measurement and structural model components while accounting for uncertainty of factor-covariance estimates: (a) plausible values and (b) the Structural-After-Measurement (SAM) approach. We conduct population analyses by varying levels of construct reliability and numbers of indicators per factor, under populations with simple and complex measurement models. Results show SAM is as promising as existing decomposed SFIs. Plausible values provide less accurate estimates, but future research should investigate whether its pooled test statistic has nominal Type I error rates.

Keywords Structural equation modeling · Construct reliability · Plausible values · Structural-after-measurement · Goodness-of-fit

G. G. Rifenbark (✉)
University of Connecticut, Storrs, CT, USA
e-mail: graham.rifenbark@uconn.edu

T. D. Jorgensen
University of Amsterdam, Amsterdam, The Netherlands

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_14

1 Evaluating Structural Fit

A structural equation model (SEM) can include both measurement and structural components. The *measurement model* pertains to the relationship between observed and latent variables (i.e., shared variance among indicators of a common factor, vs. error variance unique to each indicator). The *structural model* represents the theorized causal structure among latent variables. Evaluating how well a hypothesized SEM is substantiated by data can be conducted by (a) a null-hypothesis (H_0) test of exact fit, using the likelihood-ratio test (LRT or χ^2) statistic, or (b) quantifying approximate (mis)fit using at least one global fit index (GFI), such as the root-mean-squared error of approximation (RMSEA) or comparative fit index (CFI; see Hu & Bentler, 1998, for an overview).

When the goal is to test/evaluate the hypothesized structural model, its evaluation is complicated by qualities of the measurement model. Specifically, greater construct reliability (determined by the magnitude of factor loadings and the number of indicators per factor in the measurement model) manifests worse apparent data–model fit (e.g., higher χ^2 or RMSEA, lower CFI). That is, the same structural misspecification is easier to detect when using instruments with larger loadings or more indicators than when using fewer or less reliable indicators. Hancock and Mueller (2011) refer to this as the *reliability paradox*: lower reliability yields better apparent data–model fit, inadvertently motivating researchers to use poor-quality measurement instruments. Two existing methods for assessing structural-model fit are conflated and decomposed approaches.

Conflated approaches attempt to examine structural model fit by keeping the SEM intact, estimating both components simultaneously. A single SEM's χ^2 statistic conflates misspecification from both components, so Anderson and Gerbing (1988) proposed evaluating structural-model fit with a LRT by comparing a SEM (with hypothesized structural restrictions) to an unrestricted confirmatory factor analysis (CFA), on the assumption[1] that misspecification can only occur in the measurement component. Exact fit is thus tested with a $\Delta\chi^2_{\Delta df}$ statistic: the difference between the hypothesized SEM's χ^2_H and the structurally saturated CFA's χ^2_S, with $\Delta df = df_H - df_S$. Approximate structural fit can be evaluated using this $\Delta\chi^2$ statistic (and Δdf) in place of a single SEM's χ^2 statistic (and df) when calculating common GFIs, for example:

$$\text{RMSEA}_{(\text{D})}(\text{or RDR}) = \frac{(\Delta)\chi^2 - (\Delta)df}{(\Delta)df \times N}. \tag{1}$$

When using $\Delta\chi^2_{\Delta df}$, Browne and Du Toit (1992) referred to Eq. (1) as the root-deterioration per restriction (RDR), which Savalei et al. (2023) more recently called

[1] The structural component might be misspecified even in a CFA if the number of factors is incorrect (Mulaik & Millsap, 2000).

$RMSEA_D$. In the specific context of comparing a CFA to a structurally restricted SEM, McDonald and Ho (2002) called it RMSEA-Path, which is the term we use throughout this chapter.

Incremental fit indices (e.g., CFI) can also be calculated using $\Delta\chi^2_{\Delta df}$ (Savalei et al., 2023), but must also include the χ^2_0 statistic for a structural "null" model—e.g., an independence model with endogenous factors orthogonal to themselves and to exogenous factors—which must be nested in the hypothesized SEM (and CFA). Like Savalei et al. (2023) did with RMSEA, Lance et al. (2016) unified some past definitions by proposing a family of structural fit indices (SFIs) called "C9" that are analogous to incremental GFIs, as well as their complement (C10 = 1− C9) that quantifies badness rather than goodness of fit. For example, a C9 analogous to the normed fit index (NFI; Bentler & Bonett, 1980) is:

$$C9 = \frac{\chi^2_0 - \chi^2_H}{\chi^2_0 - \chi^2_S}, \tag{2}$$

$$C10 = \frac{\chi^2_S - \chi^2_H}{\chi^2_0 - \chi^2_S}. \tag{3}$$

One can replace each model's χ^2 in Eq. (2) with estimated noncentrality parameter (NCP) $\chi^2 - df$ for a C9 analogous to CFI, or with the ratio $\frac{\chi^2}{df}$ for a C9 analogous to the nonnormed fit index (NNFI; Bentler & Bonett, 1980) or Tucker–Lewis (1973) index (TLI).

Conversely, *decomposed* approaches examine structural model fit by separately estimating the measurement and structural components of a SEM in two steps. First, an unrestricted CFA is fitted and its model-implied latent covariance matrix ($\widehat{\boldsymbol{\Phi}}$) is extracted. Second, $\widehat{\boldsymbol{\Phi}}$ is used as input data for subsequent path analysis that models the hypothesized relations among latent variables (i.e., matching the target SEM's structural component). Two-stage estimation attempts to circumvent the reliability paradox by removing the (Stage-1) measurement model's influence on (Stage-2) structural model. Hancock and Mueller (2011) proposed calculating GFIs for the Stage-2 path analysis to serve as SFIs.

1.1 Issues with Current Methods

Conflated SFIs have nominal Type-I error rates under correct specification (Lance et al., 2016; Rifenbark, 2019, 2022), but their power to detect structural misspecification is moderated by the magnitude of factor loadings (McNeish & Hancock, 2018). Thus conflated C9/C10 still suffer the reliability paradox: C9 indicates better fit with smaller than larger factor loadings.

Although the decomposed approach appears to disentangle measurement-model misfit from structural misspecifications (Hancock & Mueller, 2011), their SFIs also suffer from inflated Type-I error rates (Rifenbark, 2022; Heene et al., 2021) when rule-of-thumb cutoffs are used (e.g., Hu & Bentler, 1999). Imprecision when

estimating $\widehat{\boldsymbol{\Phi}}$ increases an SFI's sampling variance, which occurs when measuring the factors less reliably (lower factor loadings, fewer indicators). This broadening of an SFI's sampling distribution sends more values past the "critical value" (cutoff), even when a structural model is correctly specified (i.e., due to sampling error alone; Marsh et al., 2004).

Ideally, one would not use fixed cutoffs to judge the quality of a model with SFIs (Groskurth et al., 2021; McNeish & Wolf, 2023); however, while it remains common practice, it is valuable to investigate the practical consequences of doing so. Hancock and Mueller (2011) did not propose a decomposed H_0 test of exact fit because treating the Stage-1 $\widehat{\boldsymbol{\Phi}}$ as observed data would inflate the Type I error rate. Thus, only approximate-fit solutions have been proposed from a decomposed perspective.

1.2 Potential Remedies for Evaluating Structural Fit

An ideal method would allow structural misspecifications to be identified independent from measurement-model misfit, but without ignoring the measurement model's imprecision when using $\widehat{\boldsymbol{\Phi}}$ as input data. A true test of exact fit with nominal Type I error rate would also be welcome.

We explore two potential solutions based on factor score regression (FSR; Thurstone, 1935; Thomson, 1934), which uses factor-score estimates (derived from Stage-1 measurement models) as input data for a path analysis. FSR suffers from the same limitation as decomposed SFIs: the input data are estimated (not known) factor scores, whose imprecision is not accounted for in Stage-2 estimation. One solution is numerical, the other is analytical.

1.2.1 Numerical Solution: Sample Plausible Values

Rather than obtain a single point estimate of subject i's vector of factor scores, we can draw a sample of *plausible values* from their sampling distribution, whose variance reflects their imprecision. It was first proposed for Item Response Theory (IRT; Mislevy et al., 1992; von Davier et al., 2009) and has since been applied in SEM (Asparouhov & Muthén, 2010; Jorgensen et al., 2022). The motivation is similar to sampling multiple imputations of missing values (Rubin, 1987), where the (100%-)missing values are the factor scores. Drawing m samples of plausible values provides m imputed data sets, where M should be large enough to minimize additional Monte Carlo sampling error.

To use plausible values to evaluate a structural model's fit, we first estimate an unrestricted CFA, draw m samples of plausible values, fit the hypothesized structural model (as a path analysis) to each of the m data sets, then use Rubin's (1987) rules to pool parameter estimates across m results. The LRT statistic can also be pooled

(Meng & Rubin, 1992) and the pooled statistic can be used to calculate SFIs in Eqs. 1 and 2. Variability of results across m imputations (i.e., between-imputation variance) captures the uncertainty around $\widehat{\boldsymbol{\Phi}}$ and factor scores estimated from it. Imprecision should therefore be accounted for, resulting in decomposed SFIs that yield more robust inferences about structural fit, including a test of exact fit with approximately nominal Type I error rate.

1.2.2 Analytical Solution: Use Bias-Correcting Formulas

Croon (2002) developed a bias-correcting method for FSR, which Devlieger et al. (2016) showed outperforms other FSR methods in terms of bias, mean-squared error, and Type I error rates. Devlieger et al. (2019) extended Croon's (2002) correction to construct fit indices (RMSEA, CFI, SRMR) and approximate χ^2 for nested-model tests, validating their method with simulation results. These analytical solutions even outperform SEM when there are fewer observations than indicators.

More recently, Rosseel and Loh (2022) developed *structural-after-measurement* (SAM) which generalizes Croon's correction further to be applicable when analyzing summary statistics ($\widehat{\boldsymbol{\Phi}}$) rather than raw data. Thus, factor-score estimates are no longer required. SAM is implemented in the R package `lavaan` (Rosseel, 2012) via the `sam()` function. As the name implies, measurement parameters are estimated first, potentially in separate independent measurement blocks to prevent misfit from propagating across factors (e.g., cross-loadings, residual correlations between indicators of different factors). There can be as many measurement blocks as there are latent variables or as few as one, and there are equivalent "local" and "global" SAM procedures (Rosseel & Loh, 2022). Only local SAM provides a "pseudo-χ^2 statistic" (and fit indices calculated with it) to evaluate the fit of the structural model, so we focus only on local SAM.

2 Asymptotic Investigation

We compared how well SFIs from SAM or plausible values could evaluate structural fit, relative to the flawed decomposed SFIs (Hancock & Mueller, 2011) and to the conflated test (Anderson & Gerbing, 1988) and SFIs (Lance et al., 2016). We analyze population moments at the factor level ($\boldsymbol{\Phi}$) and item level ($\boldsymbol{\Sigma}$) to obtain asymptotic results free from sampling error. Factor-level results enable us to determine "true" values (benchmarks for SFIs) of an overly restricted structural model. Item-level results enable evaluating how much each method's SFIs are affected by different measurement-model conditions.

2.1 Hypotheses

We know from past research (McNeish & Hancock, 2018) that for a given structural misspecification, SFIs of Lance et al. (2016) indicate better (or worse) fit with lower (or higher) factor loadings and fewer (or more) indicators; conversely, Hancock and Mueller (2011) SFIs are not affected (on average) by measurement quality. However, measurement-model misspecifications (e.g., omitted cross-loadings) should bias estimates of factor (co)variances, thus biasing even Hancock and Mueller (2011) SFIs.

Regardless of whether a measurement model is correctly specified, we expect plausible values to yield asymptotically identical SFIs as the decomposed SFIs of Hancock and Mueller (2011) regardless of measurement quality. Plausible values and decomposed SFIs both estimate $\widehat{\boldsymbol{\Phi}}$ from a CFA, which will not be biased by poor measurement quality, but can be biased by measurement misspecifications (e.g., omitted cross-loadings). The advantage of plausible values is that beyond SFIs, a pooled χ^2 statistic can be calculated, which should be similar to the χ^2 obtained by fitting the same model to the population $\boldsymbol{\Phi}$.

Likewise, we expect SAM to yield asymptotically identical SFIs as the decomposed SFIs of Hancock and Mueller (2011) regardless of measurement quality, but only when a measurement model is correctly specified. Given measurement misspecifications (e.g., omitted cross-loadings), SAM's independent measurement blocks provide a layer of protection from propagated errors, which should make SAM's SFIs more robust than plausible values or Hancock and Mueller (2011) SFIs.

2.2 Factor-Level Population Model

First, we specified population parameters to derive $\boldsymbol{\Phi}$, which enabled us to determine population-level SFI values for more-restricted models. We refer to these true-value results to evaluate the accuracy of SFI estimates under four different methods in the indicator level analysis. We selected a frequently used structural model for our population (Lance et al., 2016; McNeish & Hancock, 2018; Rifenbark, 2019, 2022), depicted in Fig. 1. These population parameters imply population covariance matrix $\boldsymbol{\Phi} = (\mathbf{I} - \mathbf{B})^{-1} \times \boldsymbol{\Psi} \times \left[(\mathbf{I} - \mathbf{B})^{-1}\right]'$, to which we fit four models:

- saturated Model S: all variables freely covary
- null Model 0: only X1, X2, and X3 freely covary
- true partial-mediation Model T: all paths in Fig. 1 estimated
- misspecified full-mediation Model M: Model T with fixed $\beta_{51} = \beta_{52} = 0$

Models were estimated with maximum likelihood (ML) in `lavaan()`, and the `fitMeasures()` function was used to obtain χ^2 (with $N = 500$) and GFIs for Model M. We used Models M, S, and 0 to calculate C9 (Eq. (2)), and we verified that Model T estimates matched population parameters in Fig. 1.

Model M's $\chi^2_{df=3} = 216.99$, so population RMSEA = 0.378 indicated very poor fit. Population CFI = 0.863 and analogous C9 = 0.852 were also unacceptable by most standards (Bentler & Bonett, 1980; Hu & Bentler, 1999). These "true values" are the benchmarks we will use to compare the four methods for evaluating structural fit using indicator-level data.

2.3 Indicator-Level Population Model

Holding the structural model constant, we specified different measurement models to investigate the impact of different measurement-model attributes on structural model evaluation. We manipulated three factors:

- We used 3 or 6 indicators per factor (pF). Therefore, the full SEM (Lance et al., 2016) or CFA (plausible values Hancock & Mueller, 2011) was fitted to 15 or 30 indicators. Local SAM's fitted 5 single-indicator CFAs to each factor's 3 or 6 indicators before fitting the structural component (Model M).
- Whereas McNeish and Hancock (2018) manipulated factor loadings directly, we selected loadings that would yield low or high construct reliability (CR = 0.6 or 0.9), which also depends on pF (Gagne & Hancock, 2006). As such, for a given construct reliability, factor loadings were lower when pF = 6 than when pF = 3. Table 1 shows the population $\mathbf{\Lambda}$ values (of all pF indicators) for each factor under various conditions. They are standardized loadings, such that residual variances were set to $\text{diag}(\mathbf{\Theta}) = 1 - \text{diag}(\mathbf{\Lambda\Phi\Lambda}')$.

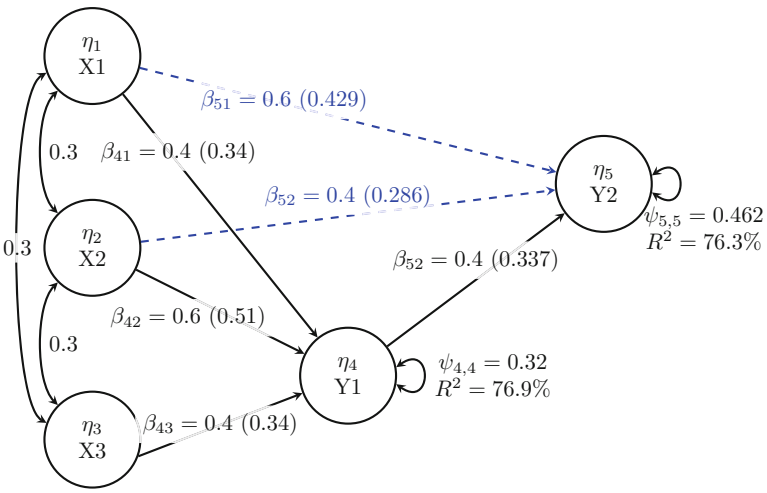


Fig. 1 Population structural parameters. Each exogenous-factor variance $\psi_{X,X} = 1$, so exogenous covariances are correlations. Standardized slopes in parentheses

– In the population, the measurement model had either simple or complex structure. Simple structure implies each observed indicator loads onto only one latent variable, and residuals are uncorrelated. Our complex measurement model contained both a cross-loading and a correlated residual. In the complex population, the covariance between the first indicators of Y1 and Y2 was $r = 0.20$ (scaled to a covariance by multiplying residual SDs: $0.2\sqrt{\theta_{y_1}\theta_{y_7}}$), and the last indicator of X3 cross-loaded onto X2. Table 1 shows that across pF and CR conditions, the cross-loading (in parentheses) was half as large as the primary loading, while maintaining indicator variances $\theta_{x,x} = 1$.

In all six conditions, we computed the population indicator-level covariance matrix implied by our SEM parameters in Fig. 1 and Table 1: $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$.

2.4 Procedure

The same four structural models that we fitted to $\boldsymbol{\Phi}$ were augmented with a simple-structure model. Thus, augmented Model S was an unrestricted CFA, augmented Model 0 was an orthogonal CFA, and augmented Models T and M were "full" SEMs representing partial and full mediation, respectively. In simple-structure conditions, the measurement model was correctly specified, but it was misspecified in complex-structure conditions because it omitted the cross-loading and residual covariance. Misspecifying the measurement model (which biases $\widehat{\boldsymbol{\Phi}}$) allowed us to compare how SFIs are influenced across the four methods.

The four full SEMs were fitted to the indicator-level population $\boldsymbol{\Sigma}$, and resulting χ^2 values were used to calculate conflated SFIs for augmented Model M: RMSEA-Path (Eq. (1); McDonald & Ho, 2002) and C9 with NCP (Eq. (2), analogous to CFI; Lance et al., 2016). To calculate decomposed versions of these SFIs (Hancock & Mueller, 2011), we saved the model-implied $\widehat{\boldsymbol{\Phi}}$ and fitted the (nonaugmented) Model M to it, just as we did to obtain "true" population SFIs by fitting Model M to the population $\boldsymbol{\Phi}$. However, $\widehat{\boldsymbol{\Phi}}$ could vary across the 2 (simple vs. complex) $\times$ 2 (pf = 3 or 6) $\times$ 2 (CR = 0.60 or 0.90) = 8 conditions.

Table 1 Population values for $\boldsymbol{\Lambda}$

	pF = 3		pF = 6	
	CR = 0.90	CR = 0.60	CR = 0.90	CR = 0.60
X1–X3	0.866	0.578	0.775	0.448
PL (CL)	0.696 (0.348)	0.464 (0.232)	0.622 (0.311)	0.359 (0.179)
Y1	0.736	0.491	0.658	0.380
Y2	0.620	0.413	0.554	0.320

Note: Simple-structure parameters given in the top row. Second row shows PL = primary loading and CL = cross-loading of indicators of X1–X3 in complex-structure conditions. Bottom rows show loadings for Y1 and Y2 under either simple or complex structure

To obtain SFIs using plausible values and SAM, raw data were necessary for analysis. We used the `rockchalk::mvrnorm()` function to generate a single data set with the argument `empirical=TRUE` to guarantee our sample's covariance matrix was identical to the population $\boldsymbol{\Sigma}$. This minimized sampling error, although some Monte Carlo error was still expected because different raw data (even with identical covariance matrices) yield different factor-score estimates.

2.4.1 Plausible Values

We fitted an unrestricted CFA (augmented Model *S*) to the raw data, then used the `semTools::plausibleValues()` function (Jorgensen et al., 2022) to sample $m = 100$ sets of plausible values. We used the `semTools::sem.mi()` function to fit Model *M* to each sample of plausible values. The `fitMeasures()` function provided SFIs using the pooled χ^2 statistic (the "D3" method; Meng & Rubin, 1992).

2.4.2 SAM

We used the `lavaan::sam()` function to fit augmented Model *M* to the raw data, which internally fitted five single-factor CFAs (i.e., 5 measurement blocks using the argument `mm=5`), followed by fitting Model *M* to the $\widehat{\boldsymbol{\Phi}}$ estimate obtained via the local-SAM method (Rosseel & Loh, 2022). SFIs are printed by the `summary()` function.

2.5 Results and Discussion

We verified that all GFIs, SFIs, and χ^2 showed perfect data–model fit when both the measurement and structural (Model *T*) components were correctly specified. Table 2 presents estimated SFIs (RMSEA and CFI) for Model *M* across conditions, with their true values from Sect. 2.2 in the column headers.

2.5.1 Conflated SFIs

As expected (McNeish & Hancock, 2018; Rifenbark, 2019, 2022), RMSEA-Path (McDonald & Ho, 2002) and C9 (Lance et al., 2016) in the $\widehat{\boldsymbol{\Sigma}}$ column of Table 2 were affected by measurement quality (CR), with lower CR inducing better apparent fit. One might not even reject the model using SFIs when construct reliability was low. Even the additional misfit from the measurement model (complex populations) did not yield SFIs that indicated fit being as poor as the true values did, although the impact of measurement misspecification was small. Holding CR constant, number

Table 2 Asymptotic estimates of SFIs across conditions

			RMSEA (= 0.378)				CFI (= 0.863)			
Measurement	CR	pF	$\widehat{\boldsymbol{\Sigma}}$	$\widehat{\boldsymbol{\Phi}}$	PV	SAM	$\widehat{\boldsymbol{\Sigma}}$	$\widehat{\boldsymbol{\Phi}}$	PV	SAM
Simple (correctly specified)	low	3	0.088	0.378	0.285	0.378	0.971	0.863	0.848	0.863
		6	0.090	0.378	0.291	0.378	0.970	0.863	0.842	0.863
	high	3	0.255	0.378	0.320	0.378	0.906	0.863	0.850	0.863
		6	0.257	0.378	0.323	0.378	0.905	0.863	0.848	0.863
Complex (misspecified)	low	3	0.080	0.356	0.254	0.358	0.977	0.890	0.888	0.887
		6	0.085	0.364	0.261	0.366	0.973	0.878	0.877	0.877
	high	3	0.251	0.373	0.309	0.373	0.910	0.872	0.865	0.871
		6	0.254	0.375	0.310	0.375	0.907	0.867	0.862	0.867

Note: True RMSEA and CFI provided in column headers as benchmarks. *CR* high (0.9) or low (0.6) construct reliability, *pF* = number of indicators per factor, $\widehat{\boldsymbol{\Sigma}}$ conflated SFIs (i.e., RMSEA-Path or C9), $\widehat{\boldsymbol{\Phi}}$ decomposed SFIs of Hancock and Mueller (2011). *PV* decomposed SFIs pooled from plausible values, *SAM* decomposed SFIs from pseudo-χ^2 of SAM approach

of indicators (pF) also did not substantially affect expected values of RMSEA-Path or C9.

2.5.2 Decomposed SFIs

When the measurement model was correctly specified, Hancock and Mueller (2011) SFIs (in the $\widehat{\boldsymbol{\Phi}}$ column of Table 2) nearly matched SAM's results across all CR and pF conditions, indicating their SFIs have asymptotically equivalent expected values. Both methods estimated true SFIs accurately for simple-structure populations. But their equivalence did not hold for misspecified measurement models. Failing to model the cross-loading and residual correlation induced small differences between SAM and Hancock and Mueller (2011) SFIs, with SAM estimates being slightly closer to true values. Although the impact of pF was small (somewhat better fit with fewer indicators), its effect was greater when CR was low.

Using plausible values also showed some promise, although its pooled SFIs were less accurate estimates of true values than SAM or Hancock and Mueller (2011). Pooled RMSEA showed better fit than the true values (particularly with low CR), and pooled CFI estimates were somewhat more accurate than RMSEA. However, pooled CFI showed better fit than true values (like RMSEA) only when the measurement model was misspecified; with correct specification, pooled CFI always showed worse fit than true values across conditions. Pooled SFIs always showed slightly worse fit with more indicators, but again this was negligible.

3 Conclusion

Population analyses show that SAM and the decomposed SFIs of Hancock and Mueller (2011) are identical in the case of the simple measurement model. However, slight differences were observed when the complex measurement model was misspecified. This was expected because SAM isolates local misfit in each measurement block, which may enable SAM to outperform Hancock and Mueller (2011) in cases of greater measurement misspecification.

In the current investigation, Hancock and Mueller (2011) SFIs appear asymptotically equivalent to SAM's SFIs. Although their sampling distributions may have the same expected values, their sampling variances may yet differ. Caution is warranted until Monte Carlo studies reveal whether increasing either's sampling variability inflates Type I errors (i.e., in smaller samples and lower CR). Holding CR constant, pF had negligible impact on SFIs, which warrants ignoring it in future Monte Carlo study, varying only CR via the magnitude of factor loadings.

The oddly inconsistent plausible-value results are likely due to the relative misfit of Model 0 and Model M, but could also be due to Monte Carlo sampling error (we drew a finite sample of plausible values, so these results were not entirely asymptotic). Results could also depend on the method for pooling the χ^2 statistic; alternatives include the "D2" method (Li et al., 1991) and "D4" (Chan & Meng, 2017). Grund et al. (2021) found that D2 can be too liberal, while D3 and D4 can be too conservative. Given how these patterns could be exacerbated in the extremely poor-fitting null Model 0, further investigation is warranted. The greatest promise of plausible values may not be for SFIs themselves, but in its ability to provide an actual (pooled) *test* of the H_0 of exact fit.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*(3), 411–423.

Asparouhov, T., & Muthén, B. (2010). Plausible values for latent variables using M*plus*. Available from http://www.statmodel.com/download/Plausible.pdf.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588–606.

Browne, M. W., & Du Toit, S. H. (1992). Automated fitting of nonstandard models. *Multivariate Behavioral Research, 27*(2), 269–300.

Chan, K. W., & Meng, X.-L. (2017). Multiple improvements of multiple imputation likelihood ratio tests. *Statistica Sinica, 32*, 1489–1514.

Croon, M. (2002). Using predicted latent scores in general latent structure models. In G. Marcoulides, & I. Moustaki (Eds.), *Latent variable and latent structure models* (pp. 195–223). Erlbaum.

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement, 76*(5), 741–770.

Devlieger, I., Talloen, W., & Rosseel, Y. (2019). New developments in factor score regression: Fit indices and a model comparison test. *Educational and Psychological Measurement, 79*(6), 1017–1037.

Gagne, P., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research, 41*(1), 65–83.

Groskurth, K., Bluemke, M., & Lechner, C. (2021). Why we need to abandon fixed cutoffs for goodness-of-fit indices: A comprehensive simulation and possible solutions. Available from PsyArXiv: https://doi.org/10.31234/osf.io/5qag3.

Grund, S., Lüdtke, O., & Robitzsch, A. (2021). Pooling methods for likelihood ratio tests in multiply imputed data sets. Available at PsyArXiv: https://doi.org/10.31234/osf.io/d459g.

Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement, 71*(2), 306–324.

Heene, M., Maraun, M. D., Glushko, N. J., & Pornprasertmanit, S. (2021). The devil is mainly in the nuisance parameters: Performance of structural fit indices under misspecified structural models in SEM. Available on PsyArXiv: https://doi.org/10.31234/osf.io/d8tuy.

Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424–453.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). `semTools`: *Useful tools for structural equation modeling*. R package version 0.5-6.

Lance, C. E., Beck, S. S., Fan, Y., & Carter, N. T. (2016). A taxonomy of path-related goodness-of-fit indices and recommended criterion values. *Psychological Methods, 21*(3), 388–404.

Li, K.-H., Meng, X.-L., Raghunathan, T. E., & Rubin, D. B. (1991). Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica, 1*(1), 65–92. Retrieved from https://www.jstor.org/stable/24303994.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*(3), 320–341.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*(1), 64–82.

McNeish, D., & Hancock, G. R. (2018). The effect of measurement quality on targeted structural model fit indices: A comment on lance, beck, fan, and carter (2016). *Psychological Methods, 23*(1), 184–190.

McNeish, D., & Wolf, M. G. (2023). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods, 28*(1), 61–88. https://doi.org/10.1037/met0000425

Meng, X.-L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika, 79*(1), 103–111.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Chapter 3: Scaling procedures in NAEP. *Journal of Educational Statistics, 17*(2), 131–154.

Mulaik, S. A., & Millsap, R. E. (2000). Doing the four-step right. *Structural Equation Modeling, 7*(1), 36–73.

Rifenbark, G. G. (2019). *Misfit at the intersection of measurement quality and model size: A Monte Carlo examination of methods for detecting structural model misspecification*. Ph.D Thesis, University of Connecticut.

Rifenbark, G. G. (2022). Impact of construct reliability on proposed measures of structural fit when detecting group differences: A Monte Carlo examination. In Wiberg, M., Molenaar, D., González, J., Kim, J.-S., & Hwang, H. (Eds.), *Quantitative psychology: The 86th annual meeting of the psychometric society, Virtual, 2021* (pp. 313–328). Springer.

Rosseel, Y. (2012). `lavaan`: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

Rosseel, Y., & Loh, W. W. (2022). A structural after measurement approach to structural equation modeling. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000503

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

Savalei, V., Brace, J. C., & Fouladi, R. T. (2023). We need to change how we compute RMSEA for nested model comparisons in structural equation modeling. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000537

Thomson, G. H. (1934). The meaning of "i" in the estimate of "g". *British Journal of Psychology. General Section, 25*(1), 92–99.

Thurstone, L. L. (1935). *The vectors of mind: Multiple-factor analysis for the isolation of primary traits*. University of Chicago Press.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1–10.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. In M. von Davier, & D. Hastedt (Eds.), *IERI monograph series: Issues and methodologies in large-scale assessments* (pp. 9–36). IEA-ETS Research Institute.

Clustering Individuals Based on Multivariate EMA Time-Series Data

Mandani Ntekouli, Gerasimos Spanakis, Lourens Waldorp, and Anne Roefs

Abstract In the field of psychopathology, Ecological Momentary Assessment (EMA) methodological advancements have offered new opportunities to collect time-intensive, repeated and intra-individual measurements. This way, a large amount of data has become available, providing the means for further exploring mental disorders. Consequently, advanced machine learning (ML) methods are needed to understand data characteristics and uncover hidden and meaningful relationships regarding the underlying complex psychological processes. Among other uses, ML facilitates the identification of similar patterns in data of different individuals through clustering. This paper focuses on clustering multivariate time-series (MTS) data of individuals into several groups. Since clustering is an unsupervised problem, it is challenging to assess whether the resulting grouping is successful. Thus, we investigate different clustering methods based on different distance measures and assess them for the stability and quality of the derived clusters. These clustering steps are illustrated on a real-world EMA dataset, including 33 individuals and 15 variables. Through evaluation, the results of kernel-based clustering methods appear promising to identify meaningful groups in the data. So, efficient representations of EMA data play an important role in clustering.

This study is part of the project "New Science of Mental Disorders" (www.nsmd.eu), supported by the Dutch Research Council and the Dutch Ministry of Education, Culture and Science (NWO gravitation grant number 024.004.016).

M. Ntekouli (✉) · G. Spanakis
Department of Advanced Computing Sciences, Maastricht University, Maastricht, The Netherlands
e-mail: m.ntekouli@maastrichtuniversity.nl; jerry.spanakis@maastrichtuniversity.nl

L. Waldorp
Department of Psychological Methods, University of Amsterdam, Amsterdam, The Netherlands
e-mail: L.J.Waldorp@uva.nl

A. Roefs
Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands
e-mail: a.roefs@maastrichtuniversity.nl

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_15

Keywords Ecological momentary assessment · EMA · Time-series data · Clustering · Cluster stability · Silhouette coefficient · DTW · Global alignment kernel

1 Introduction

In the course of EMA studies, time-intensive, repeated and intra-individual measurements are collected through digital questionnaires and smartphone's app logs and sensors. Recent methodological advancements in collecting EMA data have offered new opportunities to collect a large amount of data on a personalized level, both in terms of time points and different variables of interest. Having more time points is always a desirable data characteristic, but when more variables are involved, training a linear Vector Autoregressive (VAR) model becomes computationally expensive, and sometimes even not feasible. Especially in a complex field as psychopathology, behaviors and psychological processes are prone to interact in a non-linear fashion. Thus, applying more complex and non-linear models becomes necessary.

Such complex models can be borrowed from the field of Machine Learning (ML). ML includes a wide range of advanced statistical and probabilistic techniques that learn to build models based on the provided data (Han et al., 2022). As a result, those models are able to uncover hidden characteristics and patterns in data. A popular example is through unsupervised clustering analysis. One application of clustering in EMA data can be to identify similar individuals (Genolini et al., 2016). Although all individuals exhibit their own characteristics, they may share common influences that lead to some similar behavior. So, information of people belonging to similar groups could potentially improve the baseline personalized models (Ntekouli et al., 2022).

This paper focuses on clustering multivariate time-series (MTS) data of different individuals into several groups. For clustering time-series data, various decisions should be made regarding the clustering algorithm, distance metric and the optimal number of clusters. Thus, the most efficient methods for these decisions are described in great detail. Finally, it is proposed that validation is performed through intrinsic methods examining quality and stability of clusters. This is an important part of this paper, given that validation of time-series clustering is considered as the most challenging part.

2 Background on EMA Time-Series Data Characteristics

Before describing the clustering process, an introduction to EMA time-series' characteristics is necessary. A key point, as well as a challenge of the current problem, is the multi-level structure of EMA data. During an EMA study, data are collected sequentially, at fixed time-intervals for all participating individuals. An

example could be every 2 h for a period of 2–4 weeks. As a result, the captured data represent different aspects of participants' emotions over time and other contextual information.

When observing such a dataset, more special characteristics appear and need to be taken into account. First, some measurements can be missing, mostly because of a machine or human error. This leads to datasets with incomplete time-series. Missing points affect also the time intervals between two consecutive measurements. When missing points exist, data are characterized as irregularly spaced MTS. In such cases, beyond deletion and imputation strategies, there are still ways to process data with missing values without relying on possibly biased techniques. A widely proposed approach is to apply a kernel to the raw data. Kernel methods have dominated ML because of their effectiveness in dealing with a variety of learning problems. To tackle these problems, a kernel can be applied to map data to a reproducing kernel Hilbert space (RKHS), that is higher dimension feature space. The success of kernel methods relies on the fact that nonlinear data structures, like high dimensional MTS, can be transformed based on the type of kernel to a space where they are finally linearly separable.

Apart from length invariances, resulting from missing values, EMA time-series data can also exhibit different characteristics in terms of measurement scale and shift invariances. Regarding scaling, although EMA responses are usually recorded on a Likert scale, where 5 or 7 categories are available, the range of given responses may differ per participant. For example, some individuals may tend to be biased towards the middle values, avoiding all the extreme scores, whereas others may do the opposite, resulting in a higher skewness in some items, like negative emotions. In such cases, data normalization or scaling is a useful approach, whose effect is shown in Fig. 1b.

Additionally, different individuals' time-series can exhibit shift invariances. Time-series represent the evolution of individual's emotion or behavior. Thus, among different individuals, similar patterns of a behavior can be seen shifted in

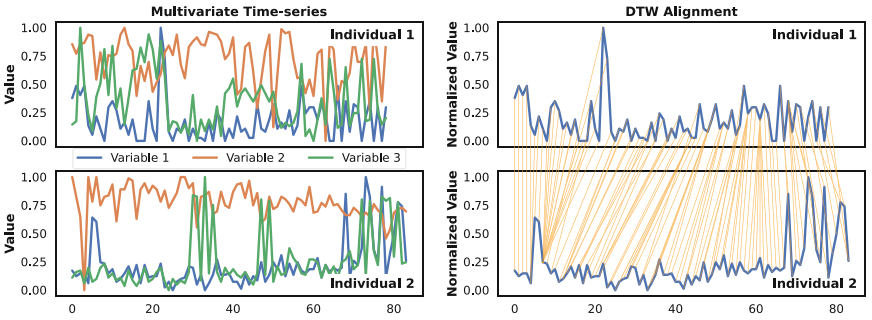


Fig. 1 (**a**) An example of 3 variables over time for 2 different individuals. (**b**) Best alignment between 2 individuals according to all variables. For the illustration, only Variable 1 is shown

time. To be able to identify these shifted patterns and consider them as similar, an appropriate alignment method should be applied. For instance, alignment issues can be taken into account by an appropriate distance measure such as DTW, that will be further discussed later.

Before applying clustering, all the aforementioned special characteristics of time-series should be taken into account (Paparrizos and Gravano, 2015). Thus, preprocessing and efficient data representations are required as additional steps.

3 Clustering Methodological Steps

In this section, an overview of all the necessary steps and decisions for applying an EMA clustering is given. We examine all the decisions regarding distance metrics and clustering methods as well as how clustering options and results can be efficiently evaluated (Von Luxburg et al., 2010).

3.1 Distance Metric

Clustering algorithms are always relying on finding the most similar elements of a dataset and group them together. Similarity can be estimated by various distance metrics, each one reflecting a different characteristic, such as intensity or shape. In order to pick an adequate distance measure, the data variances, described before, have to be considered, otherwise, different clustering methods applied on the same dataset, can produce different results.

The most commonly used distance metric is the Euclidean distance, which can be used for both, tabular data and time series. A necessary requirement is that the different time-series should be of the same length. However, in the case of EMA datasets, this requirement is usually not satisfied because of missing values. A difference in the amount of missing values occurring in the data representing various individuals make the MTS to be of variant lengths.

To tackle this issue, another distance metric is widely used, Dynamic Time Warping (DTW). DTW has become the state-of-the-art distance metric because of its high accuracy and its applicability in case of variable-length time-series (Sakoe and Chiba, 1978; Javed et al., 2020). Compared to Euclidean distance, DTW takes into account the shape difference of time-series. By stretching or compressing time series along the time axis, DTW aims to find the best shape-based alignment of these. This way, it also accounts for differences between points' time interval due to missing values, but at the same time, outliers or noise do not significantly affect it. In practice, this is possible by comparing all possible alignment paths and finally get the one leading to the minimum distance. An example of the best alignment between the same EMA item of two individuals is illustrated in Fig. 1b. The vertical lines

indicate the best alignment, showing that the two time series may not be "warped" one by one.

Any distance metric can be viewed as a kernel as long as it is also positive definite (Cuturi, 2011). Due to DTW's success, it was first considered as a good candidate for a kernel, however that's not directly possible, since it is based on the Euclidean distance, which does not satisfy all the properties of a positive definite kernel (conditional positive definite). Hence, an alternative version for a time-series kernel was created which is called global alignment kernel (GAK) (Cuturi, 2011). More specifically, as GAK was based on softDTW (Cuturi and Blondel, 2017), it takes advantage of the distance score values found across all possible alignment paths, rather than the optimal path found by DTW. According to this perspective, two time-series are considered similar not only if they have at least one alignment with high score, but quite more efficient alignment paths.

3.2 Clustering Methods

Due to the heterogeneity of clustering methods, this paper is limited to representative-based algorithms. These are distance-based methods whose goal is to retrieve a number of clusters defined by some representative elements or objects, named cluster centers. Clustering methods can be divided into two main categories, hard and fuzzy clustering (Aghabozorgi et al., 2015; Javed et al., 2020; Özkoç, 2020). In hard clustering methods, such as k-means and hierarchical clustering (HC), each individual is assigned to one cluster based on the highest similarity to clusters' center. Two challenges arise: how to integrate the appropriate distance metric and how to calculate the centroid of a cluster in case it is needed.

Nevertheless, from a theoretical point of view, in the field of psychopathology, a hard clustering algorithm could not always be the most appropriate choice. Knowing that psychopathology is a dynamically evolving, rather than a fixed, health condition, makes the approach of allowing individuals belonging to different clusters a more realistic scenario. Since clusters can capture dynamics in different time periods, individuals might be better represented by more than one cluster. Furthermore, the fact that comorbidities, meaning the co-occurrence of many mental disorders, is prevalent in a high degree leads to shared psychological processes or behaviors among patients with different diagnoses (Roefs et al., 2022). Thus, clustering algorithms permitting individuals not to be strictly assigned to only one group are considered more plausible. This can be achieved by applying fuzzy clustering algorithms, such as Fuzzy c-means (FCM) and Fuzzy k-medoids (FKM).

3.3 Clustering Evaluation

A “good" clustering result is one that identifies the “optimal" number of clusters and also how good objects, or individuals in this case, are grouped into clusters. Investigating how “good" a clustering result is can be quite challenging, since usually, there are no ground truth labels (as in supervised tasks) to compare against. To overcome this issue, an intrinsic evaluation is performed.

Ad-hoc intrinsic evaluation methods assign scores to a clustering result based on cohesion and separation. Some popular methods are Inertia, Silhouette Coefficient and Davies-Bouldin Index (Han et al., 2022). Out of these, Silhouette coefficient is picked as a metric, since it takes into account both intra-cluster and inter-cluster similarities. It compares the average similarity across individuals of the same cluster to the points belonging to the closest one. To find the closest cluster, similarities among all individuals in a cluster is taken into account. Thus, it’s quite straightforward to interpret the clustering results. Its values range from -1 to 1, where 1 and -1 indicate the best and the worst clustering, respectively, whereas 0 show a meaningless grouping, for example, when similarity differences between clusters are negligible. On the other hand, in case of fuzzy clustering, additional evaluation measures have been widely adopted, further assessing the membership degree of each individual into different groups (Cuturi & Blondel, 2016). The most common ones are Partition Coefficient (PC), Partition Entropy (PE) and Xie-Beni (XB) index, all examining the fuzziness of individuals in a different way. Apart from PC (ranges from 0 to 1), PE and XB are not bounded, while the optimal number of clusters is found at the highest, lowest and highest values, respectively. Consequently, these estimates give more information about the efficiency of fuzzy clustering.

Moreover, the stability of the clustering result should be taken into account. Running a clustering algorithm multiple times may lead to different results due to different initialization values. To evaluate clustering stability, it is needed to run the clustering algorithm several times and compare the matching of individuals’ cluster assignment. After checking all label permutations, the produced distance quantifies the mean cluster disagreement across all pairs of individuals. The result represents the clustering instability index (called stability by Von Luxburg et al., 2010) and its value can range from 0 (most stable) to 1 (less stable).

Furthermore, the extracted evaluation coefficients (such as Silhouette) can also be tested for their consistency by investigating their distribution across different runs of the algorithm. If the coefficients vary a lot, then that is indication of an unstable clustering.

Summarizing, there are various methods for evaluating a good clustering approach. Thus, in this paper, a good clustering is defined as a combination of some of the aforementioned methods. More specifically, the number k of clusters is primarily determined based on a high Silhouette coefficient, but this decision should be consistent to the findings of the other evaluation indexes as well. Subsequently, cluster stability requirement should also be fulfilled. Stability is examined on the

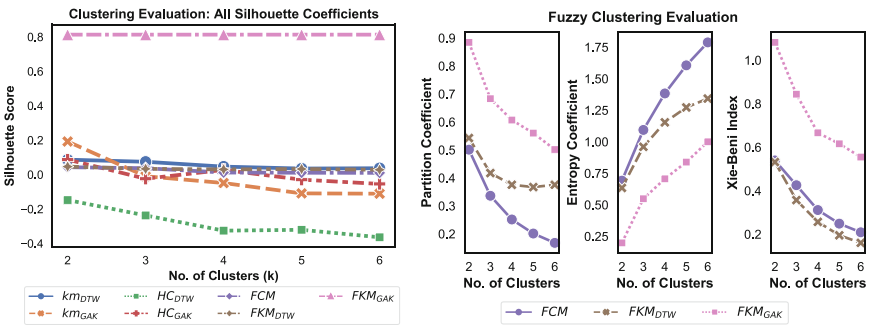
instability index as well as the consistency of silhouette coefficients when clustering is repeatedly applied.

4 Experimental Results

In this section, an example real-world dataset is used to illustrate all the decisions about methods, presented in the previous sections. The used dataset is a real-world dataset obtained by a study described in Soyster et al. (2022). It is a result of a 2-week data collection from 33 individuals, providing roughly 89 data points per individual. In a goal to capture alcohol consumption, 15 variables/indicators (such as positive and negative emotions, drinking craving and expectancies) were included in the data collection. We perform clustering on the 33 individuals based on their 15-variable time-series, taking into account the specific issues discussed in the previous chapter. Following, clustering results are evaluated through examining cluster quality and stability.

First, we apply clustering through k-means (km_{DTW}, km_{GAK}), HC (HC_{DTW}, HC_{GAK}) and FKM (FKM_{DTW}, FKM_{GAK}). Both distance metrics (DTW and GAK) are examined, except for fuzzy c-means (FCM) where only the DTW was used, as it is quite difficult to extract the clusters' centroids in the original dimensions, due to kernalization. The σ hyperparameter of the GAK kernel depends on the given data and it is calculated as the average of the median of all distances (Cuturi, 2011). Then, the groups derived by all clustering methods are evaluated in terms of Silhouette coefficient as well as stability. According to this, the optimal number of clusters is determined as well as the quality of the retrieved clusters.

Regarding the Silhouette analysis, overall results are shown in Fig. 2a. We notice that the FKM method using a GAK kernel gives the highest score. It is interesting that this remains constant for different values of clusters, as they are always grouped



(a) (b)

Fig. 2 (**a**) Maximum silhouette scores for all algorithms. (**b**) Intrinsic fuzzy clustering evaluation

to two clusters even in cases when more are allowed (leading to empty clusters). Also, a quite high score is produced by kernel k-means with $k = 2$. Apart from these, the rest of the algorithms show a result close to zero, which is interpreted as a not so meaningful clustering result. The best result among these is given by HC using a GAK kernel with $k = 2$. Therefore, it is interesting to observe that when a kernel-based method is utilized, the quality of the retrieved clusters seems to be better, showing that kernels are needed to better represent the complex structure of EMA data.

In case of fuzzy clustering methods, additional intrinsic evaluation measures can be used. The scores for different number of k are presented in Fig. 2b. These appeared to be consistent to the Silhouette results, showing that $k = 2$ is the optimal choice, also for the fuzzy clustering algorithms.

Next, we check the stability of the clustering-derived groups through silhouette scores consistency and instability index. Instability index and silhouette scores distribution were computed for 50 runs of each algorithm and are presented in Fig. 3a and b, respectively. For this part, HC is not included as it's independent of initialization issues. According to these figures, the most stable clustering result is produced by FKM, whereas the least stable by kernel k-means. A low instability score shows that groups' separation does not change a lot across repetitions. However, we can still observe an interesting case, or run, of an outlier in kernel k-means with a score approximating 0.2, which is quite higher compared to the rest. This is also apparent in Fig. 2a, for km_{GAK} and $k = 2$, and worth further investigating.

Summarizing, from a methodological perspective, various choices are possible for algorithms, distance metric and evaluation, which lead to different results. Although it is important that all methods extracted 2 clusters as the optimal grouping, it does not mean that individuals are assigned into groups in a similar way. This is also reflected when getting different results during evaluation. It is interesting to highlight that the method evaluated as the most stable is FKM_{GAK},

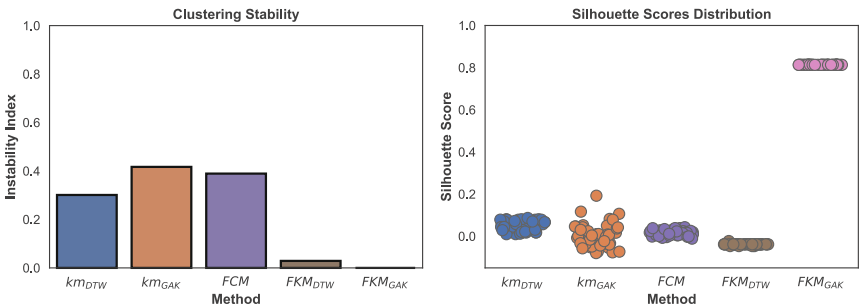


Fig. 3 Clustering evaluation for $k = 2$. (**a**) Clustering instability index. (**b**) Distributions of silhouette scores

regardless of the issue of initial parameters. Also, the fact that always two clusters were retrieved, even though more were allowed, gave more evidence for the optimal number of clusters.

5 Related Work

As already discussed, applying clustering methods to time-series data has been widely explored. Some examples of review studies are Aghabozorgi et al. (2015), Javed et al. (2020), and Özkoç (2020). Considering that all well-known clustering algorithms can be used for time-series, the challenge becomes on how to pick the right distance metric. Thus, most research studies have focused on finding a good representation of time-series similarities and integrate it to clustering algorithms.

Due to the success of the shape-based time-series clustering, other DTW-variations have been suggested, by either applying some restrictions on DTW or softening the optimal distance paths using softDTW (Cuturi & Blondel, 2017). Other studies exploring different shape-based information (Vlachos et al., 2002; Paparrizos & Gravano, 2015; Genolini et al., 2016), propose the use of the longest common subsequence (LCSS), cross-correlation and Fréchet distance, respectively.

However, most studies have handled univariate time-series data. The added value of the current paper is the multi-level structure of EMA data, including several multivariate time-series. In case of multivariate time-series, kernel-based data representations have been proposed Badiane et al. (2018). Kernels based on DTW, such as GAK, were used Cuturi and Blondel (2017). Moreover, in Mikalsen et al. (2018), another time-series cluster kernel (TCK) was proposed, based on Gaussian mixture models (GMMs).

Specifically for EMA data, only little research work has been conducted as far as clustering is concerned. In Torous et al. (2018), clustering EMA data into similar meaningful groups or clusters is proposed. However, it was not applied leaving a gap that is covered in this paper. Other than this, a different goal focusing on clustering EMA items was investigated in Hebbrecht et al. (2020). In that case, clustering was used to organize a person's symptomatology into homogeneous categories of symptoms and not for grouping different individuals like in the current paper.

6 Conclusions

This paper aims to address some of the challenges of EMA data modeling by grouping or clustering similar individuals. A detailed review of all the potential directions for applying clustering based on time-series patterns. Having described the heterogeneity of existing methods, the focus was then placed on the most challenging part of clustering, which is evaluation. A combination of several well-known ad-hoc evaluation measures was proposed, examining clustering quality

through Silhouette coefficients as well as stability. According to our analysis, kernel-based clustering methods produced the best quality clusters, showing that kernels can be useful for efficient EMA data representations. Future work can include a simulation study for evaluating clustering methods in different EMA experimental scenarios as well as further exploration of data representations using different kernels, since it plays an important role in clustering. Moreover, it should be investigated how clustering-derived groups of individuals could be further utilized. For example, an interesting approach is to train group-based models for providing more accurate predictive capabilities.

References

Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering–a decade review. *Information Systems, 53*, 16–38.

Badiane, M., O'Reilly, M., & Cunningham, P. (2018). Kernel methods for time series classification and regression. In *AICS* (pp. 54–65).

Choudhry, M. S., & Kapoor, R. (2016). Performance analysis of fuzzy c-means clustering methods for mri image segmentation. *Procedia Computer Science, 89*, 749–758.

Cuturi, M. (2011). Fast global alignment kernels. In *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 929–936).

Cuturi, M., & Blondel, M. (2017). Soft-dtw: A differentiable loss function for time-series. In *International conference on machine learning* (pp. 894–903). PMLR.

Genolini, C., Ecochard, R., Benghezal, M., Driss, T., Andrieu, S., & Subtil, F. (2016). kmlshape: an efficient method to cluster longitudinal data (time-series) according to their shapes. *Plos One, 11*(6), e0150738.

Han, J., Pei, J., & Tong, H. (2022). *Data mining: Concepts and techniques*. Morgan Kaufmann.

Hebbrecht, K., Stuivenga, M., Birkenhäger, T., Morrens, M., Fried, E., Sabbe, B., & Giltay, E. (2020). Understanding personalized dynamics to inform precision medicine: a dynamic time warp analysis of 255 depressed inpatients. *BMC Medicine, 18*(1), 1–15.

Javed, A., Lee, B. S., & Rizzo, D. M. (2020). A benchmark study on time series clustering. *Machine Learning with Applications, 1*, 100001.

Mikalsen, K. Ø., Bianchi, F. M., Soguero-Ruiz, C., & Jenssen, R. (2018). Time series cluster kernel for learning similarities between multivariate time series with missing data. *Pattern Recognition, 76*, 569–581.

Ntekouli, M., Spanakis, G., Waldorp, L., & Roefs, A. (2022). Using explainable boosting machine to compare idiographic and nomothetic approaches for ecological momentary assessment data. In *International symposium on intelligent data analysis* (pp. 199–211). Springer.

Özkoç, E. (2020). Clustering of time-series data. *Data Mining-Methods, Applications and Systems*.

Paparrizos, J., & Gravano, L. (2015). k-shape: Efficient and accurate clustering of time series. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data* (pp. 1855–1870).

Roefs, A., Fried, E. I., Kindt, M., Martijn, C., Elzinga, B., Evers, A. W., Wiers, R. W., Borsboom, D., & Jansen, A. (2022). A new science of mental disorders: Using personalised, transdiagnostic, dynamical systems to understand, model, diagnose and treat psychopathology. *Behaviour Research and Therapy, 153*, 104096.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 26*(1), 43–49.

Soyster, P. D., Ashlock, L., & Fisher, A. J. (2022). Pooled and person-specific machine learning models for predicting future alcohol consumption, craving, and wanting to drink: A demonstration of parallel utility. *Psychology of Addictive Behaviors, 36*(3), 296.

Torous, J., Larsen, M. E., Depp, C., Cosco, T. D., Barnett, I., Nock, M. K., & Firth, J. (2018). Smartphones, sensors, and machine learning to advance real-time prediction and interventions for suicide prevention: a review of current progress and next steps. *Current Psychiatry Reports, 20*(7), 1–6.

Vlachos, M., Kollios, G., & Gunopulos, D. (2002). Discovering similar multidimensional trajectories. In *Proceedings 18th international conference on data engineering* (pp. 673–684). IEEE.

Von Luxburg, U., et al. (2010). Clustering stability: An overview. *Foundations and Trends® in Machine Learning, 2*(3), 235–274.

On the Relationship Between Coefficient Alpha and Closeness Between Factors and Principal Components for the Multi-factor Model

Kentaro Hayashi and Ke-Hai Yuan

Abstract Cronbach's alpha remains very important as a measure of internal consistency in the social sciences. The Spearman-Brown formula indicates that as the number of items goes to infinity, the reliability of the composite eventually approaches one. Under proper conditions, as the lower bound of the reliability the coefficient alpha also keeps increasing with the number of items. Hayashi et al. (On coefficient alpha in high-dimensions. In: Wiberg M, Molenaar D, Gonzalez J, Bockenholt U, Kim J-S (eds) Quantitative psychology: the 85th annual meeting of the psychometric society, 2020. Springer, New York, pp 127–139, 2021) showed that under the assumption of a one-factor model, the phenomenon of the coefficient alpha approaching one as the number of items increases is closely related to the closeness between factor-analysis (FA) loadings and principal-component-analysis (PCA) loadings, and also the factor score and the principal component agreeing with each other. In this work, their partial results are extended to the case with a multi-factor model, with some extra assumptions. The new results offer another way to characterize the relationship between FA and PCA with respect to the coefficient alpha under more general conditions.

Keywords Factor analysis · Reliability · Spearman-Brown formula

1 Introduction

The coefficient alpha (Cronbach, 1951) remains very important as a measure of reliability in the social sciences. Whenever a new questionnaire is developed by psychologists, the coefficient alpha is consistently reported to demonstrate that the

K. Hayashi (✉)
University of Hawaii, Honolulu, HI, USA
e-mail: hayashik@hawaii.edu

K.-H. Yuan
University of Notre Dame, Notre Dame, IN, USA
e-mail: kyuan@nd.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_16

measure has good reliability. Moreover, the coefficient alpha itself remains an active research area in psychometrics (e.g., Sijtsma, 2009; Yuan & Bentler, 2002; Zhang & Yuan, 2016).

It is well known that, under certain conditions, as a lower bound of the reliability of the composite the coefficient alpha increases as the number of items increases. The fact has also been noted via the Spearman-Brown formula (Brown, 1910; Spearman, 1910). This implies that as the number of items goes to infinity, the reliability of the composite eventually approaches 1. Therefore, the issue of reliability is closely associated with the number of manifest variables, and we will term the issue as high dimensionality.

Regarding high dimensionality, there is another interesting phenomenon. It has been known that the results of factor analysis (FA; e.g., Lawley & Maxwell, 1971) often approach those of principal component analysis (PCA; e.g., Jolliffe, 2002), especially as the number of variables increases (Guttman, 1956; Bentler & de Leeuw, 2011; Bentler & Kano, 1990; Schneeweiss & Mathes, 1995; Schneeweiss, 1997; Kijnen, 2006). Hayashi et al. (2021) showed that the coefficient alpha approaching 1 is related to the increased closeness between FA and PCA in high dimensions under the one-factor model. Here, the closeness between FA and PCA includes the closeness with respect to both their loadings and the corresponding factor/component scores. More specifically, they showed that as the number of dimensions increases the phenomenon of the coefficient alpha approaching 1 is related to four different phenomena: (1) the closeness between the FA and PCA loadings, (2) the factor scores and the principal component scores agreeing with each other, (3) the inverse of the covariance matrix of the manifest variables becoming a diagonal matrix, assuming a FA model in the population, and (4) the communalities of the FA and PCA approaching each other.

In this work, we extend their partial results proven under the one-factor model to the model with multiple factors. More specifically, we prove that as the coefficient alpha approaches 1, the results from FA and PCA converge to each other with respect to (1) the closeness between the matrix of factor loadings and the matrix of PCA loadings as well as (2) the closeness between factors and principal components. Researchers have implicitly assumed that the use of the coefficient alpha requires the instrument to have a single factor. There are few rigorous studies connecting between the coefficient alpha and the multi-factor model. Therefore, to the best of our knowledge, this work is the first one that formally associates the coefficient alpha to the multi-factor model as well as uses it to characterize the closeness between FA and PCA.

2 Definitions and Assumptions

Suppose that there exists a p-dimensional vector of random variables, $\boldsymbol{x} = (x_1, \ldots, x_p)^T$, measuring the same construct(s). Denote the covariance matrix and the correlation matrix of $\boldsymbol{x}$ as $\boldsymbol{\Sigma}$ and $\mathbf{P}$, respectively. Then the coefficient alpha is defined as:

$$\alpha(\boldsymbol{\Sigma}) = \frac{p}{p-1}\left(1 - \frac{tr(\boldsymbol{\Sigma})}{\mathbf{1}_p^T \boldsymbol{\Sigma} \mathbf{1}_p}\right).$$

When the x_j's are all standardized, the alpha coefficient is defined as:

$$\alpha(\mathbf{P}) = \frac{p}{p-1}\left(1 - \frac{tr(\mathbf{P})}{\mathbf{1}_p^T \mathbf{P} \mathbf{1}_p}\right) = \frac{p\overline{\rho}}{1 + (p-1)\overline{\rho}},$$

where $\overline{\rho} = ({\mathbf{1}_p}^T \mathbf{P} \mathbf{1}_p - p)/p^*$ is the average of the $p^* = p(p-1)$ correlation coefficients between the distinct pairs of the items in $\boldsymbol{x}$ (Hayashi & Kamata, 2005).

The factor analysis (FA) model is expressed as $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda f} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\mu} = E(\boldsymbol{x})$ is a $p \times 1$ vector of intercepts, $\boldsymbol{\Lambda}$ is a $p \times m$ matrix of factor loadings, $\boldsymbol{f}$ is an $m \times 1$ vector of factors (latent variables), and $\boldsymbol{\varepsilon}$ is a $p \times 1$ vector of random errors. Here, we assume the number of factors m is finite (i.e., $m < \infty$). Without loss of generality, we let $\boldsymbol{\mu} = \mathbf{0}$. We assume that the mean and the variance of the factors and the errors are $E(\boldsymbol{f}) = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{f}) = \boldsymbol{I}_m$, $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, and $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}$, where $\boldsymbol{I}_m$ is an identity matrix of order m and $\boldsymbol{\Psi}$ is a diagonal matrix with positive elements. Also, we assume that the factors and the errors are uncorrelated (i.e., $\mathrm{Cov}(\boldsymbol{f}, \boldsymbol{\varepsilon}) = \mathbf{0}$). Then, the covariance matrix of $\boldsymbol{x}$ is expressed as $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$. Thus, $\boldsymbol{\Lambda}$ can be defined as $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_+ \boldsymbol{\Omega}^{1/2}$, where $\boldsymbol{\Lambda}_+$ is the standardized eigenvectors corresponding to the m nonzero eigenvalues of $\boldsymbol{\Sigma} - \boldsymbol{\Psi}$ and $\boldsymbol{\Omega}$ is the diagonal matrix whose diagonal elements are the m nonzero eigenvalues of $\boldsymbol{\Sigma} - \boldsymbol{\Psi}$ (i.e., $\boldsymbol{\Omega} = \mathrm{diag}\{\mathrm{ev}(\boldsymbol{\Sigma} - \boldsymbol{\Psi})\}$).

Likewise, we can express the PCA loadings as $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}^+ {\boldsymbol{\Omega}^*}^{1/2}$, where $\boldsymbol{\Lambda}^+$ is the $p \times m$ standardized eigenvectors corresponding to the m largest eigenvalues of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}^*$ is a diagonal matrix whose diagonal elements are the m largest eigenvalues of $\boldsymbol{\Sigma}$ (i.e., $\boldsymbol{\Omega}^* = \mathrm{diag}\{\mathrm{ev}(\boldsymbol{\Sigma})\}$). Then, the first m principal components (PCs) are obtained as $\boldsymbol{f}_+ = \boldsymbol{\Lambda}^{+T} \boldsymbol{x}$. Note $\mathrm{Cov}(\boldsymbol{f}_+) = \boldsymbol{\Lambda}^{+T} \boldsymbol{\Sigma} \boldsymbol{\Lambda}^+ = \boldsymbol{\Omega}^*$. If we define $\boldsymbol{f}^* = {\boldsymbol{\Omega}^*}^{-1/2} \boldsymbol{f}_+ = {\boldsymbol{\Omega}^*}^{-1/2} \boldsymbol{\Lambda}^{+T} \boldsymbol{x}$, then $\mathrm{Cov}(\boldsymbol{f}^*) = \boldsymbol{I}_m$. Because $\boldsymbol{\Sigma} = \sum_{i=1}^{p} \omega_i \boldsymbol{\lambda}_i^* \boldsymbol{\lambda}_i^{*T}$, where ω_i is the i-th largest eigenvalue of $\boldsymbol{\Sigma}$ and $\boldsymbol{\lambda}_i^*$ is the corresponding standardized eigenvector, we can also express $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^* \boldsymbol{\Lambda}^{*T} + \boldsymbol{\Psi}^*$, where $\boldsymbol{\Psi}^* = \sum_{i=m+1}^{p} \omega_i^* \boldsymbol{\lambda}_i^* \boldsymbol{\lambda}_i^{*T}$ is not a diagonal matrix, in general, unlike the FA model.

We employ the average squared canonical correlations between the two loading matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}^*$ (Schneeweiss & Mathes, 1995; Schneeweiss, 1997) as a measure of closeness between $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}^*$. The squared canonical correlations are given by the eigenvalues of $(\boldsymbol{\Lambda}^T \boldsymbol{\Lambda})^{-1} (\boldsymbol{\Lambda}^T \boldsymbol{\Lambda}^*) (\boldsymbol{\Lambda}^{*T} \boldsymbol{\Lambda}^*)^{-1} (\boldsymbol{\Lambda}^{*T} \boldsymbol{\Lambda})$, and are known to be invariant with respect to orthogonal rotations. Thus, the average squared canonical correlation between matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}^*$ is given by

$$\rho^2(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^*) = (1/m)\,\mathrm{tr}\left\{\left(\boldsymbol{\Lambda}^T \boldsymbol{\Lambda}\right)^{-1} \left(\boldsymbol{\Lambda}^T \boldsymbol{\Lambda}^*\right) \left(\boldsymbol{\Lambda}^{*T} \boldsymbol{\Lambda}^*\right)^{-1} \left(\boldsymbol{\Lambda}^{*T} \boldsymbol{\Lambda}\right)\right\}.$$

Note that $\rho^2(\mathbf{\Lambda}, \mathbf{\Lambda}^*)$ is well defined whenever the inverse of $\mathbf{\Lambda}^T\mathbf{\Lambda}$ and $\mathbf{\Lambda}^{*T}\mathbf{\Lambda}^*$ exist, which is true whenever $\mathbf{\Lambda}^T\mathbf{\Lambda}$ and $\mathbf{\Lambda}^{*T}\mathbf{\Lambda}^*$ are positive definite or when both $\mathbf{\Lambda}$ and $\mathbf{\Lambda}^*$ are of full column rank (Theorem 14.2.9 of Harville, 1997). A special case of the average squared canonical correlation is when the two matrices are column vectors $\boldsymbol{f}$ and $\boldsymbol{f}^*$:

$$\begin{aligned}\rho^2(\boldsymbol{f}, \boldsymbol{f}^*) &= \left(\boldsymbol{f}^T\boldsymbol{f}\right)^{-1}\left(\boldsymbol{f}^T\boldsymbol{f}^*\right)\left(\boldsymbol{f}^{*T}\boldsymbol{f}^*\right)^{-1}\left(\boldsymbol{f}^{*T}\boldsymbol{f}\right) \\ &= \left(\boldsymbol{f}^T\boldsymbol{f}^*\right)^2/\left\{||\boldsymbol{f}||_2||\boldsymbol{f}^*||_2\right\}^2,\end{aligned}$$

which is equal to the squared correlation between $\boldsymbol{f}$ and $\boldsymbol{f}^*$. Note that the correlation here is defined slightly different from that of the Pearson correlation, without centering.

Assumption 1. The diagonal elements of $\mathbf{\Sigma}$ are finite (i.e., $\sigma_{ii} \leq \sigma_{\sup} < \infty$) and the unique (error) variances (elements of $\mathbf{\Psi}$) are bounded away from zero ($0 < \psi_{\inf} \leq \psi_{ii}$).

Note: Then, the unique (error) variances are also bounded above, i.e., $0 < \psi_{\inf} \leq \psi_{ii} \leq \psi_{\sup} \leq \sigma_{\sup} < \infty$. Also, the diagonal elements (σ_{ii}) of $\mathbf{\Sigma}$ are also bounded away from zero, because $0 < \psi_{\inf} \leq \sigma_{\inf} \leq \sigma_{ii}$.

Assumption 2. The average correlation $\overline{\rho}$ is positive, bounded away from zero, and strictly less than 1 (i.e., $0 < c \leq \overline{\rho} < 1$ for some small $c > 0$).

Assumption 3. The sum of squared elements of $\mathbf{\Lambda}$ is of order p.

Note: The sum of squared elements of $\mathbf{\Lambda}$ is equal to the sum of diagonal elements (i.e., the sum of m eigenvalues) of $\mathbf{\Lambda}^T\mathbf{\Lambda}$. So, we can express *Assumption* 3 as $\mathrm{tr}(\mathbf{\Lambda}^T\mathbf{\Lambda}) = (C)(p)$ with some $C < \infty$.

3 Theorem

(1) If the coefficient alpha approaches 1 (i.e., $\alpha(\mathbf{\Sigma}) \to 1$), then the matrix of FA loadings ($\mathbf{\Lambda}$) and the matrix of PCA loadings ($\mathbf{\Lambda}^*$) converge to each other with respect to the average squared canonical correlation (i.e., $\rho^2(\mathbf{\Lambda}, \mathbf{\Lambda}^*) \to 1$).
(2) If the coefficient alpha approaches 1 (i.e., $\alpha(\mathbf{\Sigma}) \to 1$), then the factors ($\boldsymbol{f}$) and the principal components ($\boldsymbol{f}^*$) converge to each other with respect to the average squared correlation (i.e., $\rho^2(\boldsymbol{f}, \boldsymbol{f}^*) \to 1$).

4 Lemmas and Proof of the Theorem

Lemma 1. $\alpha(\mathbf{\Sigma}) \to 1$ as $p \to \infty$ if and only if $\alpha(\mathbf{P}) \to 1$ as $p \to \infty$.
Note: *Lemma* 1 implies that in proving the *Theorem*, we can work with the alpha coefficient in either the metric of the covariance matrix or that of the correlation matrix.

Proof: Let $\mathbf{\Sigma} = \mathbf{\Delta P \Delta}$, where $\mathbf{\Delta} = \text{diag}(\sigma_{ii}{}^{1/2})$ is the diagonal matrix whose diagonal elements are the standard deviations. Due to *Assumption* 1, the diagonal elements (σ_{ii}) of $\mathbf{\Sigma}$ are bounded away from zero, and we can express $\mathbf{P}$ as $\mathbf{P} = \mathbf{\Delta}^{-1}\mathbf{\Sigma}\mathbf{\Delta}^{-1}$. Thus, there exist

$$\mathbf{0} < \sigma_{\text{inf}}\mathbf{P} \le \mathbf{\Sigma} \le \sigma_{\text{sup}}\mathbf{P} < \infty \text{and } \mathbf{0} < \sigma_{\text{sup}}{}^{-1}\mathbf{\Sigma} \le \mathbf{P} \le \sigma_{\text{inf}}{}^{-1}\mathbf{\Sigma} < \infty.$$

(<=) Taking the trace on each term of the inequality $\sigma_{\text{inf}}\mathbf{P} \le \mathbf{\Sigma} \le \sigma_{\text{sup}}\mathbf{P}$ yields $(\sigma_{\text{inf}})\text{tr}(\mathbf{P}) \le \text{tr}(\mathbf{\Sigma}) \le (\sigma_{\text{sup}})\text{tr}(\mathbf{P})$. It follows from $\sigma_{\text{inf}}{}^{1/2}\boldsymbol{I}_p \le \mathbf{\Delta} \le \sigma_{\text{sup}}{}^{1/2}\boldsymbol{I}_p$ that $(\sigma_{\text{inf}})(\mathbf{1}_{\text{p}}{}^{\text{T}}\mathbf{P}\mathbf{1}_{\text{p}}) \le \mathbf{1}_p{}^T\mathbf{\Sigma}\mathbf{1}_p = \mathbf{1}_p{}^T\mathbf{\Delta P \Delta}\mathbf{1}_p \le (\sigma_{\text{sup}})(\mathbf{1}_p{}^T\mathbf{P}\mathbf{1}_p)$. Thus,

$$\begin{aligned}\alpha(\mathbf{\Sigma}) &= \{p/(p-1)\}\left\{1-\text{tr}(\mathbf{\Sigma}) / \left(\mathbf{1}_p{}^T\mathbf{\Sigma}\mathbf{1}_p\right)\right\} \\ &\ge \{p/(p-1)\}\left\{1-\left(\sigma_{\text{sup}}\right)\text{tr}(\mathbf{P}) / \left[(\sigma_{\text{inf}})\,\mathbf{1}_p{}^T\mathbf{P}\mathbf{1}_p\right]\right\} \\ &= \{p/(p-1)\}\left\{1-\left(\sigma_{\text{sup}}/\sigma_{\text{inf}}\right)\text{tr}(\mathbf{P}) / \left(\mathbf{1}_p{}^T\mathbf{P}\mathbf{1}_p\right)\right\}.\end{aligned}$$

Now, $p/(p-1) \to 1$ as $p \to \infty$, and $\alpha(\mathbf{P}) = \{p/(p-1)\}\{1 - \text{tr}(\mathbf{P})/(\mathbf{1}_p{}^T\mathbf{P}\mathbf{1}_p)\} \to 1$ implies $\text{tr}(\mathbf{P})/(\mathbf{1}_p{}^T\mathbf{P}\mathbf{1}_p) \to 0$. Because σ_{sup} is bounded and σ_{inf} is bounded away from zero, $\sigma_{\text{sup}}/\sigma_{\text{inf}}$ is also bounded and bounded away from zero. Thus, $\text{tr}(\mathbf{P})/(\mathbf{1}_p{}^T\mathbf{P}\mathbf{1}_p) \to 0$ implies $(\sigma_{\text{sup}}/\sigma_{\text{inf}})\text{tr}(\mathbf{P})/(\mathbf{1}_p{}^T\mathbf{P}\mathbf{1}_p) \to 0$, and $\alpha(\mathbf{\Sigma}) \to 1$ follows as $p \to \infty$ (with $p/(p-1) \to 1$).
(=>) Taking the trace on each term of the inequality $\sigma_{\text{sup}}{}^{-1}\mathbf{\Sigma} \le \mathbf{P} \le \sigma_{\text{inf}}{}^{-1}\mathbf{\Sigma}$ yields $(\sigma_{\text{sup}}{}^{-1})\text{tr}(\mathbf{\Sigma}) \le \text{tr}(\mathbf{P}) \le (\sigma_{\text{inf}}{}^{-1})\text{tr}(\mathbf{\Sigma})$. Also, $(\sigma_{\text{sup}}{}^{-1})(\mathbf{1}_p{}^T\mathbf{\Sigma}\mathbf{1}_p) \le \mathbf{1}_p{}^T\mathbf{P}\mathbf{1}_p = \mathbf{1}_p{}^T\mathbf{\Delta}^{-1}\mathbf{\Sigma}\mathbf{\Delta}^{-1}\mathbf{1}_p \le (\sigma_{\text{inf}}{}^{-1})(\mathbf{1}_p{}^T\mathbf{\Sigma}\mathbf{1}_p)$ follows from $\sigma_{\text{sup}}{}^{-1}\mathbf{\Sigma} \le \mathbf{P} \le \sigma_{\text{inf}}{}^{-1}\mathbf{\Sigma}$. Thus,

$$\begin{aligned}\alpha(\mathbf{P}) &= \{p/(p-1)\}\left\{1-\text{tr}(\mathbf{P}) / \left(\mathbf{1}_p{}^T\mathbf{P}\mathbf{1}_p\right)\right\} \\ &\ge \{p/(p-1)\}\left\{1-\left(\sigma_{\text{inf}}{}^{-1}\right)\text{tr}(\mathbf{\Sigma}) / \left(\left(\sigma_{\text{sup}}{}^{-1}\right)\mathbf{1}_p{}^T\mathbf{\Sigma}\mathbf{1}_p\right)\right\} \\ &= \{p/(p-1)\}\left\{1-\left(\sigma_{\text{sup}}/\sigma_{\text{inf}}\right)\text{tr}(\mathbf{\Sigma}) / \left(\mathbf{1}_p{}^T\mathbf{\Sigma}\mathbf{1}_p\right)\right\}.\end{aligned}$$

As before, because $(\sigma_{\text{sup}}/\sigma_{\text{inf}})$ is bounded and bounded away from zero, $\text{tr}(\mathbf{\Sigma})/(\mathbf{1}_p{}^T\mathbf{\Sigma}\mathbf{1}_p) \to 0$ implies $(\sigma_{\text{sup}}/\sigma_{\text{inf}})\text{tr}(\mathbf{\Sigma})/(\mathbf{1}_p{}^T\mathbf{\Sigma}\mathbf{1}_p) \to 0$. Thus $\alpha(\mathbf{P}) \to 1$ follows as $p \to \infty$ (with $p/(p-1) \to 1$). □

Lemma 2. If α ($\mathbf{P}$) $\to 1$ then $p \to \infty$.

Proof: $\alpha(\mathbf{P}) = \{1 + p^{-1}(\overline{\rho}^{-1} - 1)\}^{-1} \to 1$ is equivalent to $p^{-1}(\overline{\rho}^{-1} - 1) \to 0$. With *Assumption* 2 ($0 < c \leq \overline{\rho} < 1$), noting $0 < \overline{\rho}^{-1} - 1 \leq c^{-1} - 1 < \infty$, $p^{-1}(\overline{\rho}^{-1} - 1) \to 0$ implies $p^{-1} \to 0$ (i.e., $p \to \infty$).

Lemma 3 (Schneeweiss, 1997, Theorem 1 (1)).

If the smallest (the m-th) eigenvalue of $\mathbf{\Lambda}^T \mathbf{\Psi}^{-1} \mathbf{\Lambda} \to \infty$ (i.e., if $\mathrm{ev}_m(\mathbf{\Lambda}^T \mathbf{\Psi}^{-1} \mathbf{\Lambda}) \to \infty$) as $p \to \infty$, then $\rho^2(\mathbf{\Lambda}, \mathbf{\Lambda}^*) \to 1$.

Proof: Because we modify the original proof by Schneeweiss (1997), we give a full proof. Due to *Assumption* 1 ($0 < \psi_{\inf} \leq \psi_{ii} \leq \psi_{\sup} \leq \sigma_{\sup} < \infty$), $\mathrm{ev}_m(\mathbf{\Lambda}^T \mathbf{\Psi}^{-1} \mathbf{\Lambda}) \to \infty$ implies $\mathrm{ev}_m(\mathbf{\Lambda}^T \mathbf{\Lambda}) \to \infty$. Also, note that, because $\mathbf{\Lambda}^* = \mathbf{\Lambda}^+ \mathbf{\Omega}^{*1/2}$, $\mathbf{\Lambda}^{*T} \mathbf{\Lambda}^* = \mathbf{\Omega}^{*1/2} \mathbf{\Lambda}^{+T} \mathbf{\Lambda}^+ \mathbf{\Omega}^{*1/2} = \mathbf{\Omega}^*$.

Now, noting the well-known property of the determinant ($\det(\boldsymbol{AB}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$) for square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, let $\boldsymbol{R} = (\mathbf{\Lambda}^T \mathbf{\Lambda})^{-1}(\mathbf{\Lambda}^T \mathbf{\Lambda}^*)(\mathbf{\Lambda}^{*T} \mathbf{\Lambda}^*)^{-1}(\mathbf{\Lambda}^{*T} \mathbf{\Lambda})$ and take the determinant of both sides:

$$
\begin{aligned}
\det(\boldsymbol{R}) &= \det\left(\mathbf{\Lambda}^T \mathbf{\Lambda}^* \mathbf{\Lambda}^{*T} \mathbf{\Lambda}\right) / \left\{\det\left(\mathbf{\Lambda}^T \mathbf{\Lambda}\right) \det\left(\mathbf{\Lambda}^{*T} \mathbf{\Lambda}^*\right)\right\} \\
&= \det\left(\mathbf{\Lambda}^{*T} \mathbf{\Lambda} \mathbf{\Lambda}^T \mathbf{\Lambda}^*\right) / \left\{\det\left(\mathbf{\Lambda}^T \mathbf{\Lambda}\right) \det\left(\mathbf{\Omega}^*\right)\right\} \\
&= \det\left(\mathbf{\Lambda}^{*T} (\mathbf{\Sigma} - \mathbf{\Psi}) \mathbf{\Lambda}^*\right) / \left\{\det\left(\mathbf{\Lambda}^T \mathbf{\Lambda}\right) \det\left(\mathbf{\Omega}^*\right)\right\} \\
&= \det\left(\mathbf{\Omega}^{*2} - \mathbf{\Lambda}^{*T} \mathbf{\Psi} \mathbf{\Lambda}^*\right) / \left\{\det\left(\mathbf{\Lambda}^T \mathbf{\Lambda}\right) \det\left(\mathbf{\Omega}^*\right)\right\} \\
&\geq \det\left(\mathbf{\Omega}^{*2} - \psi_{\sup} \mathbf{\Omega}^*\right) / \left\{\det\left(\mathbf{\Lambda}^T \mathbf{\Lambda}\right) \det\left(\mathbf{\Omega}^*\right)\right\} \\
&= \det\left(\mathbf{\Omega}^* - \psi_{\sup} \boldsymbol{I}_m\right) / \det\left(\mathbf{\Lambda}^T \mathbf{\Lambda}\right).
\end{aligned}
$$

Let ω^* be the smallest diagonal element of $\mathbf{\Omega}^* = \mathrm{diag}(\omega_j^*)$. Then $\omega^* = \mathrm{ev}_m(\mathbf{\Sigma}) = \mathrm{ev}_m(\mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}) \geq \mathrm{ev}_m(\mathbf{\Lambda}\mathbf{\Lambda}^T) = \mathrm{ev}_m(\mathbf{\Lambda}^T \mathbf{\Lambda}) \to \infty$ as $p \to \infty$. Let $\boldsymbol{A}(p)$ be a finite-dimensional square matrix which is a function of p. Then, because the determinant $\det: \boldsymbol{M}_{n \times n}(\mathbb{R}) \to \mathbb{R}$ is a continuous function, if $\lim_{p \to \infty} \boldsymbol{A}(p)$ exists, we can always interchange the limit with the determinant (i.e., $\lim_{p \to \infty} \det(\boldsymbol{A}(p)) = \det(\lim_{p \to \infty} \boldsymbol{A}(p))$) (See e.g., https://math.stackexchange.com/questions/3684644/can-we-interchange-the-limit-with-determinant). Now, define two $\boldsymbol{A}(p)$'s as $\boldsymbol{A}_1(p) = (\mathbf{\Omega}^* - \psi_{\sup} \boldsymbol{I}_m) / \omega^*$ and $\boldsymbol{A}_2(p) = \mathbf{\Lambda}^T \mathbf{\Lambda} / \omega^*$. We have $\lim_{p \to \infty} \det(\boldsymbol{A}_1(p)) = \det(\lim_{p \to \infty} \boldsymbol{A}_1(p))$ and $\lim_{p \to \infty} \det(\boldsymbol{A}_2(p)) = \det(\lim_{p \to \infty} \boldsymbol{A}_2(p))$. Here, we assumed that the limits of both $\prod_{j=1}^m \left(\omega_j^* / \omega^*\right)$ and $\det(\mathbf{\Lambda}^T \mathbf{\Lambda} / \omega^*)$ exist. Thus, we have

$$
\begin{aligned}
\lim_{p\to\infty} \det(\boldsymbol{R}) &\geq \lim_{p\to\infty} \left\{\det\left(\boldsymbol{\Omega}^{*} - \psi_{\sup}\boldsymbol{I}_m\right) / \det\left(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}\right)\right\} \\
&= \lim_{p\to\infty} \frac{\prod_{j=1}^{m}\left(\omega_j^{*} - \psi_{\sup}\right)}{\det\left(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}\right)} \\
&= \lim_{p\to\infty} \frac{\prod_{j=1}^{m}\left(\omega_j^{*}/\omega^{*} - \psi_{\sup}/\omega^{*}\right)}{\det\left(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}/\omega^{*}\right)} \\
&= \lim_{p\to\infty} \frac{\prod_{j=1}^{m}\left(\omega_j^{*}/\omega^{*}\right)}{\det\left(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}/\omega^{*}\right)} \\
&= \lim_{p\to\infty} \frac{\prod_{j=1}^{m}\omega_j^{*}}{\det\left(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}\right)} \\
&= \lim_{p\to\infty} \left\{\det\left(\boldsymbol{\Omega}^{*}\right) / \det\left(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}\right)\right\} \\
&\geq 1 = \det\left(\boldsymbol{I}_{\mathrm{m}}\right).
\end{aligned}
$$

Therefore, $\det(\boldsymbol{R}) \to \det(\boldsymbol{I}_m)$, and $\boldsymbol{R} \to \boldsymbol{I}_m$ follows. Thus $\rho^2(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^{*}) = (1/m)\mathrm{tr}(\boldsymbol{R}) \to 1$ follows.

Now, we remain to prove that $\det(\boldsymbol{R}) \to \det(\boldsymbol{I}_m) = 1$ implies $\boldsymbol{R} \to \boldsymbol{I}_m$. Suppose $\boldsymbol{R} \to \boldsymbol{A}$ and $\det(\boldsymbol{A}) = 1$ but $\boldsymbol{A}$ is not an identity matrix. Then, there exist at least one eigenvalue of $\boldsymbol{A}$ that is not equal to 1, which, due to $\det(\boldsymbol{A}) = \prod_{i=1}^{m} \mathrm{ev}_i(\boldsymbol{A}) = 1$, implies that there must be an eigenvalue of $\boldsymbol{A}$ greater than 1 and also an eigenvalue of $\boldsymbol{A}$ smaller than 1. However, because $\boldsymbol{R}$ is a generalized squared correlation matrix and $\boldsymbol{R} \to \boldsymbol{A}$, all the eigenvalues of $\boldsymbol{A}$ must be at most 1 (between 0 and 1).[1] Thus a contradiction results and $\boldsymbol{R} \to \boldsymbol{I}_m$ follows. □

Lemma 4 (Schneeweiss, 1997, Theorem 1 (2)). If the smallest (the m-th) eigenvalue of $\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} \to \infty$ (i.e., if $\mathrm{ev}_m(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \to \infty$), then $\rho^2(\boldsymbol{f}, \boldsymbol{f}^{*}) \to 1$.

Proof: First note $\boldsymbol{f}^{*} = \boldsymbol{\Omega}^{*-1/2}\boldsymbol{\Lambda}^{+T}\boldsymbol{x}$ and $\boldsymbol{\Lambda}^{*} = \boldsymbol{\Lambda}^{+}\boldsymbol{\Omega}^{*1/2}$, and let
$\boldsymbol{F} = \mathrm{E}(\boldsymbol{f}\,\boldsymbol{f}^{*T})\{\mathrm{E}(\boldsymbol{f}^{*}\boldsymbol{f}^{*T})\}^{-1}\mathrm{E}(\boldsymbol{f}^{*}\boldsymbol{f}^{T})\{\mathrm{E}(\boldsymbol{f}\,\boldsymbol{f}^{T})\}^{-1}$. Because $\mathrm{E}(\boldsymbol{f}\,\boldsymbol{f}^{T}) = \boldsymbol{I}_m$ and $\mathrm{E}(\boldsymbol{f}^{*}\boldsymbol{f}^{*T}) = \boldsymbol{I}_m$,
$\boldsymbol{F} = \mathrm{E}(\boldsymbol{f}\boldsymbol{f}^{*T})\mathrm{E}(\boldsymbol{f}^{*}\boldsymbol{f}^{T})$. Now, because

$$
\mathrm{E}\left(\boldsymbol{f}\;\boldsymbol{f}^{*T}\right) = \mathrm{E}\left(\boldsymbol{f}\boldsymbol{x}^T\boldsymbol{\Lambda}^{+}\boldsymbol{\Omega}^{*-1/2}\right) = \boldsymbol{\Lambda}^T\boldsymbol{\Lambda}^{+}\boldsymbol{\Omega}^{*-1/2} = \boldsymbol{\Lambda}^T\boldsymbol{\Lambda}^{*}\boldsymbol{\Omega}^{*-1} \text{ and}
$$

$$
\mathrm{E}\left(\boldsymbol{f}^{*}\boldsymbol{f}^{T}\right) = \boldsymbol{\Omega}^{*-1}\boldsymbol{\Lambda}^{*T}\boldsymbol{\Lambda},
$$

it follows that

$$
\boldsymbol{F} = \boldsymbol{\Lambda}^T\boldsymbol{\Lambda}^{*}\boldsymbol{\Omega}^{*-2}\boldsymbol{\Lambda}^{*T}\boldsymbol{\Lambda}.
$$

[1] See e.g., http://www2.tulane.edu/~PsycStat/dunlap/Psyc613/RI2.html and https://stats.stackexchange.com/questions/284861/do-the-determinants-of-covariance-and-correlation-matrices-and-or-their-inverses on this point.

Now, multiply $\boldsymbol{I}_m=(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda})^{-1}(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda})(\boldsymbol{\Lambda}^{*T}\boldsymbol{\Lambda}^*)^{-1}(\boldsymbol{\Lambda}^{*T}\boldsymbol{\Lambda}^*)$ to the right-hand side and take the determinant for the both sides. Noting $\boldsymbol{R}=(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda})^{-1}(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}^*)(\boldsymbol{\Lambda}^{*T}\boldsymbol{\Lambda}^*)^{-1}(\boldsymbol{\Lambda}^{*T}\boldsymbol{\Lambda})$, we get

$$\det(\boldsymbol{F}) = \det(\boldsymbol{R}) \det\left(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}\boldsymbol{\Omega}^{*-1}\right) \det\left(\boldsymbol{\Lambda}^{*T}\boldsymbol{\Lambda}^*\boldsymbol{\Omega}^{*-1}\right).$$

Due to *Lemma* 3, $\mathrm{ev}_m(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \to \infty$ implies $\det(\boldsymbol{R}) \to 1$ and $\rho^2(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^*) \to 1$. As shown in the proof of *Lemma* 3, $\det(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}\boldsymbol{\Omega}^{*-1}) \to 1$, and $\rho^2(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^*) \to 1$ also implies $\det(\boldsymbol{\Lambda}^{*T}\boldsymbol{\Lambda}^*\boldsymbol{\Omega}^{*-1}) \to 1$. Thus $\det(\boldsymbol{F}) \to 1$, that is, $\boldsymbol{F} \to \boldsymbol{I}_m$ and $\rho^2(\boldsymbol{f}, \boldsymbol{f}^*) \to 1$ follows.

Alternatively, we can prove *Lemma* 4 without using *Lemma* 3, as follows (See Schneeweiss & Mathes, 1995, Theorem 1). Noting that $\boldsymbol{\Lambda}^+$ is a matrix of eigenvectors, so that $\boldsymbol{\Lambda}^{+T}\boldsymbol{\Lambda}^+ = \boldsymbol{I}_{\mathrm{m}}$,

$$\begin{aligned}
\rho^2\left(\boldsymbol{f}, \boldsymbol{f}^*\right) &= (1/m)\,\mathrm{tr}\,(\boldsymbol{F}) \\
&= (1/m)\,\mathrm{tr}\left(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}^*\boldsymbol{\Omega}^{*-2}\boldsymbol{\Lambda}^{*T}\boldsymbol{\Lambda}\right) \\
&= (1/m)\,\mathrm{tr}\left(\boldsymbol{\Omega}^{*-1}\boldsymbol{\Lambda}^{*T}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}^*\boldsymbol{\Omega}^{*-1}\right) \\
&= (1/m)\,\mathrm{tr}\left(\boldsymbol{\Omega}^{*-1}\boldsymbol{\Lambda}^{*T}\,(\boldsymbol{\Sigma}-\boldsymbol{\Psi})\,\boldsymbol{\Lambda}^*\boldsymbol{\Omega}^{*-1}\right) \\
&= (1/m)\,\mathrm{tr}\left(\boldsymbol{\Omega}^{*-1/2}\boldsymbol{\Lambda}^{+T}\,(\boldsymbol{\Sigma}-\boldsymbol{\Psi})\,\boldsymbol{\Lambda}^+\boldsymbol{\Omega}^{*-1/2}\right) \\
&= (1/m)\,\mathrm{tr}\left\{\boldsymbol{\Omega}^{*-1/2}\left(\boldsymbol{\Omega}^*-\boldsymbol{\Lambda}^{+T}\boldsymbol{\Psi}\boldsymbol{\Lambda}^+\right)\boldsymbol{\Omega}^{*-1/2}\right\} \\
&= 1-(1/m)\,\mathrm{tr}\left(\boldsymbol{\Omega}^{*-1}\boldsymbol{\Lambda}^{+T}\boldsymbol{\Psi}\boldsymbol{\Lambda}^+\right) \\
&\geq 1-\left(\psi_{\sup}\right)(1/m)\,\mathrm{tr}\left(\boldsymbol{\Omega}^{*-1}\boldsymbol{\Lambda}^{+T}\boldsymbol{\Lambda}^+\right) \\
&= 1-\left(\psi_{\sup}\right)(1/m)\,\mathrm{tr}\left(\boldsymbol{\Omega}^{*-1}\right) \\
&\geq 1-\left(\psi_{\sup}\right)(1/m)\,\mathrm{tr}\left(\left(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}\right)^{-1}\right) \\
&\to 1.
\end{aligned}$$

The last inequality follows because $\boldsymbol{\Omega}^* \geq \boldsymbol{\Lambda}^T\boldsymbol{\Lambda}$, $\boldsymbol{\Omega}^{*-1} \leq (\boldsymbol{\Lambda}^T\boldsymbol{\Lambda})^{-1}$, and $-\boldsymbol{\Omega}^{*-1} \geq -(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda})^{-1}$. □

Lemma 5. $p \to \infty$ implies the smallest (the m-th) eigenvalue of $\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} \to \infty$ (i.e., If $p \to \infty$, then $ev_m(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \to \infty$).

Proof: With *Assumption* 1 and under the assumption that $p \to \infty$, $ev_m(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \to \infty$ is equivalent to $ev_m(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}) \to \infty$. Obviously, $ev_m(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}) \to \infty$ implies $ev_1(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}) \to \infty$. Also, $ev_1(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}) < \infty$ implies $ev_m(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}) < \infty$. Thus, we instead prove the statement: "If $p \to \infty$, then $ev_1(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}) \to \infty$," which is equivalent to "If $ev_1(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}) < \infty$, then $p < \infty$." Now, because for a finite m, $ev_1(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}) < \infty$ implies the sum of the m largest eigenvalues of $\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}$ is also finite, by *Assumption* 3, $\mathrm{tr}(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}) = (C)(p) < \infty$ with some $C < \infty$. Thus $p < \infty$ follows.

Now, we prove the *Theorem*.

(1) Due to *Lemma* 1, we prove the claim in the correlation metric (with $\alpha(\mathbf{P})$). First, due to *Lemma* 2, α ($\mathbf{P}$) $\rightarrow$ 1 implies $p \rightarrow \infty$. Second, due to *Lemma* 5, $p \rightarrow \infty$ implies $ev_m(\mathbf{\Lambda}^T\mathbf{\Psi}^{-1}\mathbf{\Lambda}) \rightarrow \infty$. Finally, due to *Lemma* 3, $ev_m(\mathbf{\Lambda}^T\mathbf{\Psi}^{-1}\mathbf{\Lambda}) \rightarrow \infty$ implies $\rho^2(\mathbf{\Lambda}, \mathbf{\Lambda}^*) \rightarrow 1$.
(2) Again, due to *Lemma* 1, we prove the claim in the correlation metric (with $\alpha(\mathbf{P})$). As before, due to *Lemma* 2, α ($\mathbf{P}$) $\rightarrow$ 1 implies $p \rightarrow \infty$. Next, due to *Lemma* 5, $p \rightarrow \infty$ implies $ev_m(\mathbf{\Lambda}^T\mathbf{\Psi}^{-1}\mathbf{\Lambda}) \rightarrow \infty$. (Up to this point, same as (1).) Finally, due to *Lemma* 4, $ev_m(\mathbf{\Lambda}^T\mathbf{\Psi}^{-1}\mathbf{\Lambda}) \rightarrow \infty$ implies $\rho^2(\boldsymbol{f}, \boldsymbol{f}^*) \rightarrow 1$.

5 Simulation

In the previous section, we had to impose the assumption that the average correlation $\overline{\rho}$ is bounded away from zero (i.e., $0 < c \leq \overline{\rho}$ for some small $c > 0$) for the proof of the *Theorem*. We demonstrate the importance of this assumption with a small simulation in this section. We employed two correlation structures, one satisfying the assumption and the other not satisfying the assumption. For the correlation structure in which the assumption was not satisfied, the matrix of factor loadings in the population was of the form $\mathbf{\Lambda}$ ($p \times m = 2$) $= (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$, where the j-th element of $\boldsymbol{\lambda}_1$ was given by exp(−sqrt(j)), that is, $\boldsymbol{\lambda}_1 = (0.368, 0.243, 0.177, 0.135, 0.107, 0.086, \ldots)^T$ and the j-th element of $\boldsymbol{\lambda}_2$ was given by exp($-j$), that is, $\boldsymbol{\lambda}_2 = (0.368, 0.135, 0.050, 0.018, 0.007, 0.002, \ldots)^T$. Thus, as j increases, the factor loadings decrease exponentially. Obviously, the loadings violate the stated assumption. In the other condition in which the assumption is satisfied, we used the same loadings as before but added a small constant of $c = 0.05$ to each loading. The number of variables p ranged from 102 to 300, and increased by 6 (i.e., $p = 102, 108, 114, 120, \ldots, 294, 300$). From the factor loadings, we constructed correlation matrices and generated data using the mvrnorm function in the MASS package in R version 4.1.3. The sample size was $n = 2000$ and the number of replications for each p was 5. We obtained the average of 5 replications, thus the number of data points for each condition was $(300 - 102)/6 + 1 = 34$. Then we compared the results.

Figure 1 shows the relationship between the coefficient alpha and the Fisher-z transformed average squared canonical correlation between FA loading and PCA loading matrices $\mathbf{\Lambda}$ and $\mathbf{\Lambda}^*$ (i.e., $z = (1/2)\log\{(1 + \rho(\mathbf{\Lambda}, \mathbf{\Lambda}^*))/(1 - \rho(\mathbf{\Lambda}, \mathbf{\Lambda}^*))\}$, where $\rho(\mathbf{\Lambda}, \mathbf{\Lambda}^*)$ is the square root of $\rho^2(\mathbf{\Lambda}, \mathbf{\Lambda}^*)$) under the two conditions. As Fig. 1 shows, there was a high positive correlation (0.787, p-value < 0.001) between the coefficient alpha and the Fisher-z transformed average squared canonical correlation when the assumption holds. On the other hand, there was no significant correlation (−0.189, p-value = 0.285) when the assumption does not hold. Next, Fig. 2 shows the relationship between the number of variables (p) and the coefficient alpha under the two conditions. As Fig. 2 shows, there was a positive linear relationship between p and the coefficient alpha (with a correlation of 0.989, p-value < 0.001) when the assumption holds. On the other hand, there was no significant slope between p and the coefficient alpha (with a correlation of −0.230, p-value −0.191) when the assumption does not hold.

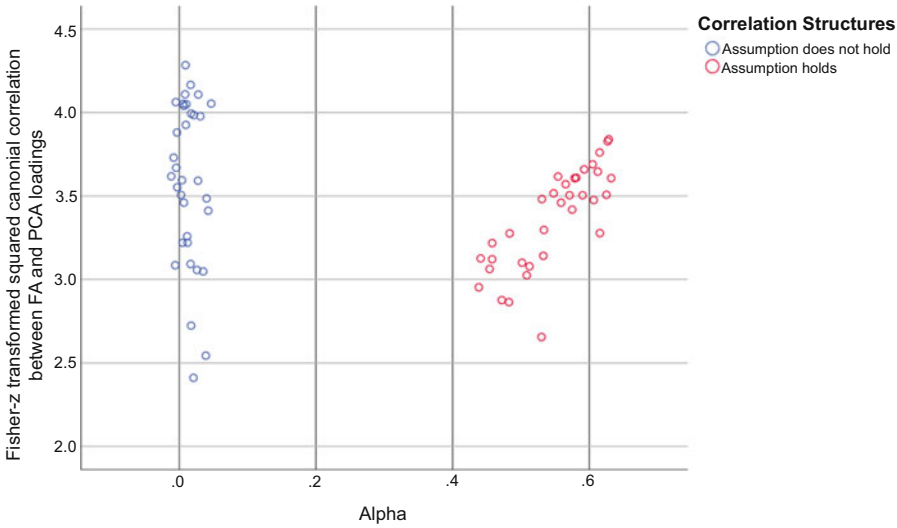


Fig. 1 Relationship between the coefficient alpha and the Fisher-z transformed average squared canonical correlation between the FA and PCA loadings

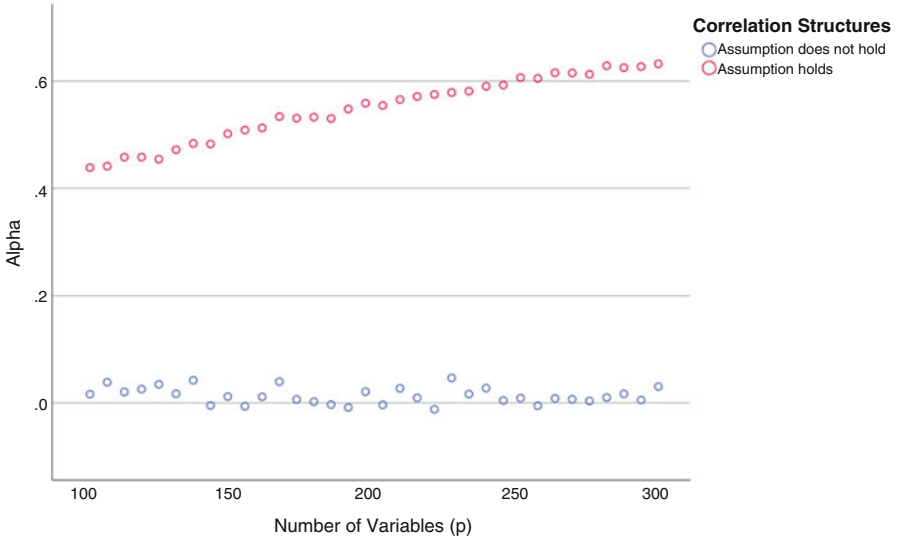


Fig. 2 The relationship between the number of variables and the coefficient alpha

6 Discussion

We showed that the phenomenon of the coefficient alpha approaching 1 is related to the increased closeness between FA and PCA for the multi-factor model. Our results are an extension of the work by Hayashi et al. (2021) who showed the

connection between the phenomenon of the coefficient alpha approaching 1 and the increased closeness between FA and PCA under the one-factor model. Our results imply that when the value of coefficient alpha is close to 1, we can use PCA as an approximation to FA and trust that the results are almost the same whether we use FA or PCA even for the multi-factor model. A practical implication for this work is that we can use the value of coefficient alpha as an index for the degree of closeness between FA and PCA.

It is well known that a set of items must follow a single-factor model in order to appropriately apply the coefficient alpha. Probably because of the strong association between the coefficient alpha and a single factor, it seems that the connection between the coefficient alpha and the multi-factor model has not been studied. To the best of our knowledge, our work is the first-ever attempt to connect the coefficient alpha to the multi-factor model.

In deriving our *Theorem*, we introduced assumptions that the average correlation $\overline{\rho}$ is positive, bounded away from zero, and strictly less than 1 (*Assumption* 2), and the sum of squared elements of factor loadings is of order p (*Assumption* 3). Alternatively, instead of these two assumptions, we can introduce the following assumptions to derive the same results stated in our *Theorem*.

Assumption 4. The proportion that factor loadings are positive and bounded away from zero ($0 < \lambda_{\text{inf}} \leq \lambda_{ij}$) approaches 1 as p increases.

Assumption 5. The average correlation is strictly less than 1 ($\overline{\rho} < 1$).

Note that *Assumption* 4 implies that, with probability 1, the average correlation $\overline{\rho}$ becomes positive and bounded away from zero, as p increases. This is because only a limited number of the off-diagonal elements ($\sigma_{ij} = \sum_{k=1}^{m} \lambda_{ik}\lambda_{jk}$, $i \neq j$) can be negative and their contribution to $\overline{\rho}$ becomes nullified as p increases. Combined with $\overline{\rho} < 1$ (*Assumption* 5), *Assumptions* 4 also implies $0 < c \leq \overline{\rho} < 1$, just like *Assumption* 2. Also, we can still prove *Lemma* 5 with *Assumption* 4 by noting the inequality

$$\infty > \text{tr}\left(\mathbf{\Lambda}^T\mathbf{\Lambda}\right) = \sum\nolimits_{j=1}^{m}\sum\nolimits_{i=1}^{p}\lambda_{ij}^2 \geq \left(\lambda_{\text{inf}}{}^2\right)(m)(p)$$

with an infimum of factor loadings λ_{inf} ($0 < \lambda_{\text{inf}} \leq \lambda_{ij}$), in place of $\text{tr}(\mathbf{\Lambda}^T\mathbf{\Lambda}) = (C)(p) < \infty$ in *Assumption* 3.

Practically speaking, the alternative assumption of positive loadings (and also a positive matrix) may not be so unnatural. First, note the signs of the factors are indeterminate. If there is a column in which all factor loadings are negative, we can reverse the sign of those factor loadings to positive without changing the correlation structure. Second, often, positive factor loadings can be found after rotations are performed even if the initial solution of the maximum likelihood estimates includes negative loadings. Third, positive matrices and positive loading matrices can be found among examples well-known to researchers. For example, two examples in

Lawley and Maxwell (1971) are both positive matrices (see Tables 4.1 and 4.4). All varimax-rotated factor loadings of the correlation matrices given in Tables 4.1 and 4.3 are positive (see the table in the middle of p. 84 and Table 6.6 on p. 76 of Lawley & Maxwell, 1971). Also, the correlation matrix given in Schneeweiss (1997) is a positive matrix, except for one off-diagonal element whose value is 0.

However, *Assumption* 4 may still be a rather strong assumption compared to *Assumptions* 2 and 3. With *Assumptions* 2 and 3, we no longer require that almost all factor loadings are positive or that almost all off-diagonal elements of the correlation matrix are positive. Consequently, there are probably numerous examples in which *Assumptions* 2 and 3 hold. For example, the correlation matrix given in Table 15.2 of Press (2003) is for 15 variables, and out of 105 unique off-diagonal elements, 4 entries are negative. So, this correlation matrix is not a positive matrix but it still satisfies *Assumptions* 2 and 3. It will not be difficult to find many similar examples. Thus, we believe that our results have wide applicability in practice with *Assumptions* 2 and 3.

Acknowledgments The authors would like to thank Dr. Dylan Molenaar for his careful review of the manuscript. This work was supported by a grant from the Department of Education (R305D210023), and by a grant from the Natural Science Foundation of China (31971029). However, the contents of the study do not necessarily represent the policy of the funding agencies, and you should not assume endorsement by the Federal Government.

References

Bentler, P. M., & de Leeuw, J. (2011). Factor analysis via component analysis. *Psychometrika, 76*, 461–470. https://doi.org/10.1007/s11336-011-9217-5

Bentler, P.M. & Kano, Y. (1990). On the equivalence of factors and components. *Multivariate Behavioral Research, 25*, 67–74. https://doi.org/10.1207/s15327906mbr2501_8

Brown, W. (1910). Some experimental results in the correlation of mental ability. *British Journal of Psychology, 3*, 271–295.

Cronbach, L. I. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. https://doi.org/10.1007/BF02310555

Guttman, L. (1956). Best possible systematic estimates of communalities. *Psychometrika, 21*, 273–285. https://doi.org/10.1007/BF02289137

Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*. Springer.

Hayashi, K., & Kamata, A. (2005). A note of the estimator of the alpha coefficient for standardized variables under normality. *Psychometrika, 70*, 579–586. https://doi.org/10.1007/s11336-001-0888-1

Hayashi, K., Yuan, K.-H., & Sato, R. (2021). On coefficient alpha in high-dimensions. In M. Wiberg, D. Molenaar, J. Gonzalez, U. Bockenholt, & J.-S. Kim (Eds.), *Quantitative psychology: The 85th annual meeting of the psychometric society, 2020* (pp. 127–139). Springer.

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). Springer.

Kijnen, W. P. (2006). Convergence of estimates of unique variances in factor analysis, based on the inverse sample covariance matrix. *Psychometrika, 71*, 193–199. https://doi.org/10.1007/s11336-000-1142-9

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method* (2nd ed.). American Elsevier.

Press, S. J. (2003). *Subjective and objective Bayesian statistics* (2nd ed.). Wiley.

Schneeweiss, H. (1997). Factors and principal components in the near spherical case. *Multivariate Behavioral Research, 32*, 375–401. https://doi.org/10.1207/s15327906mbr3204_4

Schneeweiss, H., & Mathes, H. (1995). Factor analysis and principal components. *Journal of Multivariate Analysis, 55*, 105–124. https://doi.org/10.1006/jmva.1995.1069

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107–120. https://doi.org/10.1007/s11336-008-9101-0

Spearman, C. C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Yuan, K.-H., & Bentler, P. M. (2002). On robustness of the normal-theory based asymptotic distributions of three reliability coefficient estimates. *Psychometrika, 67*, 251–259. https://doi.org/10.1007/BF02294845

Zhang, Z., & Yuan, K.-H. (2016). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: Methods and software. *Educational and Psychological Measurement, 76*, 387–411. https://doi.org/10.1177/0013164415594658

A Genetic Algorithm-Based Framework for Learning Statistical Power Manifold

Abhishek K. Umrawal, Sean P. Lane, and Erin P. Hennes

Abstract Statistical power is a measure of the replicability of a categorical hypothesis test. Formally, it is the probability of detecting an effect, if there is a true effect present in the population. Hence, optimizing statistical power as a function of some parameters of a hypothesis test is desirable. However, for most hypothesis tests, the explicit functional form of statistical power for individual model parameters is unknown; but calculating power for a given set of values of those parameters is possible using simulated experiments. These simulated experiments are usually computationally expensive. Hence, developing the entire statistical power manifold using simulations can be very time-consuming. We propose a novel genetic algorithm-based framework for learning statistical power manifolds. For a multiple linear regression F-test, we show that the proposed algorithm/framework learns the statistical power manifold much faster as compared to a brute-force approach as the number of queries to the power oracle is significantly reduced. We also show that the quality of learning the manifold improves as the number of iterations increases for the genetic algorithm. Such tools are useful for evaluating statistical power trade-offs when researchers have little information regarding a priori 'best guesses' of primary effect sizes of interest or how sampling variability in non-primary effects impacts power for primary ones.

Keywords Statistical power · Hypothesis testing · Genetic algorithm · Nearest neighbors

A. K. Umrawal (✉)
Purdue University, West Lafayette, IN, USA

University of Maryland, Baltimore County, MD, USA
e-mail: aumrawal@purdue.edu

S. P. Lane · E. P. Hennes
Purdue University, West Lafayette, IN, USA

University of Missouri, Columbia, MO, USA
e-mail: lanesp@missouri.edu; ehennes@missouri.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_17

1 Introduction

1.1 Motivation

Statistical power analysis is of great importance in empirical studies (Yang et al., 2022; Fraley & Vazire, 2014; Cafri et al., 2010). Statistical power is a measure of the goodness/strength of a hypothesis test. Formally, it is the probability of detecting an effect, if there is a true effect present to detect. For instance, if we use a test to conclude that a specific therapy or medicine is helpful in anxiety and stress alleviation then the power of the test tells us how confident we are about this insight. Hence, optimizing the statistical power as a function of some parameters of a hypothesis test is desirable. However, for most hypothesis tests, the explicit functional form of statistical power as a function of those parameters is unknown but calculating statistical power for a given set of values of those parameters is possible using simulated experiments. These simulated experiments are usually computationally expensive. Hence, developing the entire statistical power manifold using simulations can be very time-consuming. The objective of this paper is to develop a framework for learning statistical power while significantly reducing the cost of simulations.

1.2 Literature Review

Bakker et al. (2012), Bakker et al. (2016), Maxwell (2004), and Cohen (1992) point out that quite frequently the statistical power associated with empirical studies is so low that the conclusions drawn from those studies are highly unreliable. This happens primarily due to the lack of a formal statistical power analysis. Using simulations of statistical power, Bakker et al. (2012) shows that empirical studies use questionable research practices by using lower sample sizes with more trials. The results show that such practices can significantly overestimate statistical power. This makes the conclusions of the study about the effect size to be misleading as the reproducibility of the study is hampered due to the low statistical power. The questionable research practices include performing multiple trials with a very small sample size, using additional subjects before carrying out the analysis, and removal of outliers. Bakker et al. (2012) emphasizes carrying out a formal power analysis for deciding the sample size to avoid such low statistical power. Recently, Baker et al. (2021) provides an online tool for drawing power contours to understand the effect of sample size and trials per participant on statistical power. The results provided in this paper demonstrate that changes to the sample size and number of trials lead to understanding how power regions of the power manifold. Rast and Hofer (2014) also demonstrates a powerful (inversely correlated) impact of sample size on both the effect size and the study design. Recently, Lane and Hennes (2018) and Lane and Hennes (2019) provide simulation-based methods for formally conducting statistical power analysis. These simulation-based methods perform well in terms of power

estimation but can be computationally expensive due to a high number of queries to the simulation-based power function oracle.

1.3 Contribution

In this paper, we provide a novel genetic algorithm-based framework for learning the statistical power manifold in a time-efficient manner by significantly reducing the number of queries to the power function oracle. For a multiple linear regression F-test, we show that the proposed algorithm/framework learns the statistical power manifold much faster than a brute-force approach as the number of queries to the power oracle is significantly reduced. We show that the quality of learning the manifold improves as the number of iterations increases for the genetic algorithm.

2 Methodology

Let $y_1, ..., y_n$ be n sample observations from some probability distribution F with a p-dimensional parameter vector $\boldsymbol{\theta}$. Let ϕ denote the hypothesis test of interest. Let H_0 and H_1 be the corresponding *null* and *alternative* hypotheses, respectively. Let α and γ be the probabilities of Type I and Type II errors, respectively. Therefore, the power of the test is $1 - \gamma$. Let β be the effect size that is a function of the parameter vector $\boldsymbol{\theta}$ associated with the probability distribution F.

Let the sample size n be given to varying in a fixed range according to some budget constraint, and there is some expert/prior knowledge about the range in which the unknown effect size of interest may vary. Let $[\beta_l, \beta_u]$, $[\boldsymbol{\theta}_l, \boldsymbol{\theta}_u]$, and $[n_l, n_u]$ be the initial ranges of effect size, parameter vector, and sample size respectively. Statistical power associated with a hypothesis test is a function of the effect size β (hence of $\boldsymbol{\theta}$), level of significance α, and sample size n. Since the level of significance is predetermined as a fixed number, we can say that power is a function of the effect size and sample size for a fixed given value of the level of significance.

Define $\boldsymbol{c} := (\beta, n) \equiv (\boldsymbol{\theta}, n) \equiv (\theta_1, ..., \theta_p, n)$. For the hypothesis test ϕ, statistical power is a function $\boldsymbol{c}$ for a given value of α denoted as $1 - \gamma := f_{\phi,\alpha}(\boldsymbol{c})$. In most cases, we do not know the explicit algebraic form of $f(\cdot)$ to calculate power as a function of $\boldsymbol{c}$. However, calculating power for a specific choice of $\boldsymbol{c}$ is possible using simulations. For our work, we assume that we have access to such a black box that takes $\boldsymbol{c}$ as input and returns $1 - \gamma$ as output. We call this black box a *power function value oracle*.

A *brute-force* way of learning/computing the power manifold (for parameters in the initial ranges) is to divide the ranges of the parameters into a high-dimensional grid, compute the power using the power function value oracle, and then plot the values. The drawback of this approach is the computational cost associated with a very large number of queries to the value oracle.

Motivated by the idea of reducing the computational cost of learning the statistical power manifold while being able to do well in terms of power estimation, we propose a genetic algorithm-based framework.

2.1 Learning Power Manifold Using Genetic Algorithm

Genetic algorithm (Goldberg & Holland, 1988; Holland, 1992) is a meta-heuristic inspired by biological evolution based on Charles Darwin's theory of natural selection that belongs to the larger class of evolutionary algorithms. Genetic algorithms are commonly used to generate high-quality solutions to optimization and search problems by relying on biologically inspired operators such as mutation, crossover, and selection (Mitchell, 1998).

We next explain the steps involved in genetic algorithm in the context of our problem as follows.

We start with N (a hyper-parameter) randomly chosen $\boldsymbol{c}$ vectors where entries inside the vector are chosen randomly from the respective ranges with some discretization. The discretization step size for parameters $\theta_1, ..., \theta_p$ is usually a proper positive fraction and for sample size is a positive integer. Each of these N random vectors is called a *chromosome*. The collection of these N is called a *population*. In simple words, we initialize a population of N chromosomes. Let $\{\mathbf{c}_1, ..., \mathbf{c}_N\}$ be the initial population where $\mathbf{c}_i$ is the ith chromosome. A *gene* is defined as a specific entry in the chromosome vector. We next calculate the power values associated with these N chromosomes using the power function value oracle. The power value of a chromosome is called its *fitness*. Let f_i be the fitness of $\mathbf{c}_i$. We save these chromosomes and the corresponding fitness values in a hash map (dictionary) $\mathcal{D}$.

We next go to *reproduction* to form the next generation of chromosomes. The chance of *selection* of a chromosome from a past generation to the next generation is an increasing function of its fitness. The idea of using power value as fitness for reproduction is motivated by the fact that we are interested in maximizing the power and would like to travel/take steps towards the high-power region.

We next go to *mutation* where the probability of a chromosome going through mutation is p_m (a hyper-parameter). If a chromosome is selected for mutation then we mutate a randomly chosen gene of that chromosome by replacing its current value with some randomly chosen value within its initial range.

We next go to *crossover* where we select the best two chromosomes in terms of fitness. We then select a random index and split both chromosomes at that index. Finally, we merge the front piece of the first chromosome with the end piece of the second chromosome, and vice-versa.

We then *repeat* reproduction, mutation, and crossover for some I (a hyper-parameter) iterations. Note that, in these successive iterations some chromosomes are repeated. We do not need to query the power value oracle again to calculate their fitness as we save this information in a dictionary. We update this dictionary after every iteration for new (not seen previously) chromosomes.

The *final dictionary* $\mathcal{D}$ with all *(chromosome, power value)* pairs that the genetic algorithm comes across through the iterations of the genetic algorithm in the process of learning to reach a high/max power region gives us an *estimate of the power manifold* of interest.

2.2 Power Prediction Using Nearest Neighbors

So far, we have estimated the power manifold using the genetic algorithm. We next answer the following question. How do we predict the power of a new set of arguments? Note that the genetic algorithm described above comes across some sets of arguments but not all. It is indeed desirable that the above genetic algorithm queries the power value oracle as less as possible but still be able to learn the manifold well. However, in general, a user may be interested in knowing the power function value for an arbitrary set of arguments.

Once we have estimated/learned the power manifold, we can use this estimated manifold to predict the power values for an arbitrary set of arguments instead of querying the costly power value oracle. We use a simple *nearest neighbors predictor* described as follows.

For a given set of arguments, $\mathbf{c}$, select k (a hyper-parameter) nearest neighbors (in terms of the Euclidean distance) in $\mathcal{D}$ returned by the genetic algorithm. Provide a prediction of the power for $\mathbf{c}$ as the average of the power values of those k nearest neighbors in $\mathcal{D}$.

Refer to Fig. 1 for an overview of the proposed methodology.

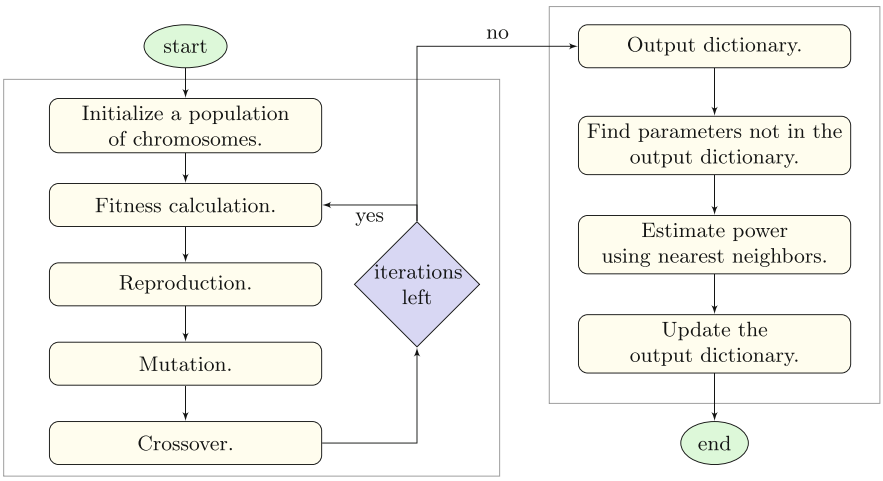


Fig. 1 The proposed methodology with the genetic algorithm on the left and the nearest neighbors on the right

3 Experiments

We perform experiments to demonstrate the performance of our algorithm in terms of power function estimation and run-time. We consider the following multiple linear regression model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 x_2) + \epsilon,$$

where y is the response, x_1 is the experimental condition (-1: control, 1: treatment), x_2 is some other measure, and $x_1 x_2$ is the interaction of the experimental condition and the other measure.

We chose the above model as it is simple to understand and also covers important aspects of experimental studies, viz., a categorical variable, a continuous variable, and an interaction of these variables.

3.1 Experimental Details

Based on similar prior knowledge we know the following. $\beta_1 \in [0.10, 0.30]$, $\beta_2 \in [0.30, 0.90]$, $\beta_3 > 0$. Based on the budget constraint, we have $n \leq 500$. The level of significance, $\alpha = 0.05$.

For genetic algorithm, we use $N = 1000, 2000, \ldots, 5000$, $I = 10, 20, \ldots, 100$, $\lambda = 1$, $p_m = 0.05$, regression coefficients' discretization step size = 0.05, sample size's discretization step size = 5, and number of simulations for the power value oracle = 1000. For nearest neighbors, we use $k = 5$. We focus on the following two tests.

1. $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.
2. $H_0 : \beta_3 = 0$ against $H_1 : \beta_3 \neq 0$.

We use a t-test to test the significance of the partial regression coefficient β_1 while controlling for the rest of the model parameters, and similarly for β_3.

3.2 Algorithms

We compute the statistical power manifold for the above setup using the proposed methodology discussed in Sect. 2. The proposed methodology first involves the genetic algorithm discussed in Sect. 2.1 and then the k-nearest neighbors discussed in Sect. 2.2.

For assessing the performance of our algorithm, we compute the power manifold using the *brute-force* method also discussed in Sect. 2.

3.3 Evaluation Metrics

We compare the performance of our algorithm with a costly brute-force strategy in terms of the quality of the estimation using root mean squared error. Let $\mathbf{c}_1, \ldots, \mathbf{c}_M$ be all set of arguments seen by the brute-force method. Let $f_1^{(b)}, \ldots, f_M^{(b)}$ be the corresponding power values computed using the brute-force method. Let $f_1^{(g)}, \ldots, f_M^{(g)}$ be the corresponding power values using our algorithm (either after the genetic algorithm or after the nearest neighbors prediction). We calculate the root mean squared error (RMSE) as follows.

$$\text{RMSE} = \left[\frac{1}{n} \sum_{i=1}^{M} \left(f_i^{(b)} - f_i^{(g)} \right)^2 \right]^{1/2}.$$

Note that RMSE for the Brute-force method will be zero.

3.4 Results

For our experiments, computing power manifold for $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$, and $H_0 : \beta_3 = 0$ against $H_1 : \beta_3 \neq 0$ using the brute-force method takes approximately 8000 s.

For different population sizes, the time taken by our algorithm as a function of the number of iterations is plotted in Fig. 2.

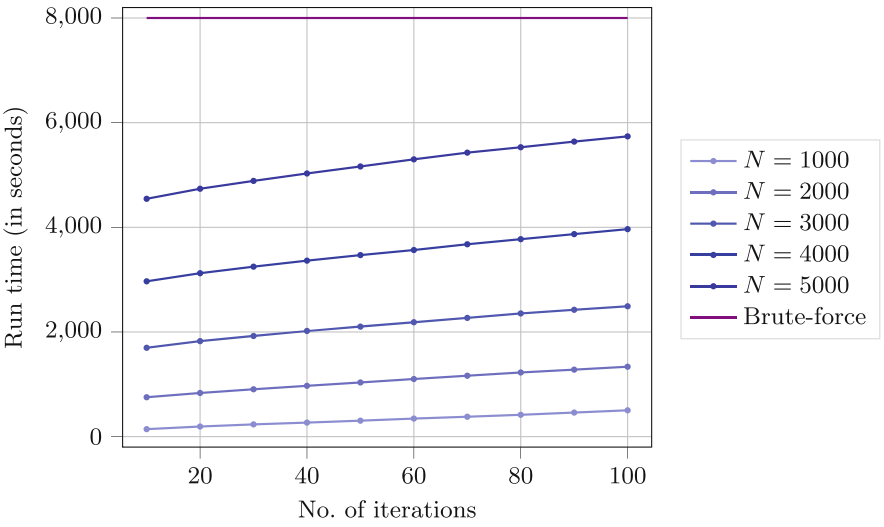


Fig. 2 Run times vs. no. of iteration. *Brute-force run-time is 8000 s*

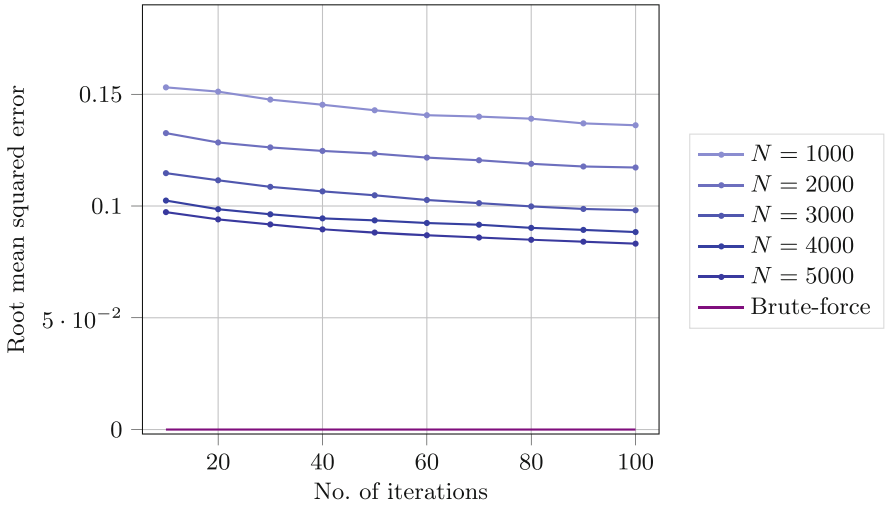


Fig. 3 Root mean squared error vs. no. of iteration. *Brute-force RMSE is zero*

For different population sizes, the root mean squared error (calculated against the brute-force estimation) for our algorithm as a function of the number of iterations is plotted in Fig. 3.

3.5 Discussion

Based on Fig. 2, we make the following observations. The time taken by our algorithm is always less than the brute-force method. The time taken by our algorithm increases as the number of iterations increases and the size of the population increases, respectively. The rate of increase in run-time as a function of the number of iterations decreases as the size of the population increases. The reason for this behavior is the following. A smaller (larger) initial population will require less (more) time in calculating the fitness of the initial population but in the future, the algorithm will come across more (less) new set of arguments.

Based on Fig. 3, we make the following observations. The root mean squared error for our algorithm decreases as the number of iterations increases and the size of the population increases, respectively. The rate of decrease in root mean squared error as a function of the number of iterations decreases as the size of the population increases.

4 Conclusion

For learning the statistical power manifold in a time-efficient manner, we developed a novel genetic algorithm-based framework. Using our algorithm, applied researchers may learn/construct the statistical power manifold for some given initial constraints on different parameters. The learned surface can be used to identify high-power and low-power regions in the power manifold that can help applied researchers design their experiments better.

We performed experiments to demonstrate the performance of the proposed algorithm. We showed that our algorithm learns the power manifold as good as the costly brute-force methods while bringing huge savings in terms of run-time. Furthermore, we showed that the quality of learning using the proposed method improves as the number of iterations of the genetic algorithm increases. However, the run-time of the proposed algorithm also increases as the number of iterations of the genetic algorithm increases. This exhibits the optimality vs. run-time trade-off associated with our algorithm. Based on the estimation threshold and the availability of computational resources, the user may choose the required number of iterations and the size of the population.

For further details about our work, refer to the technical report, Umrawal et al. (2022).

5 Future Work

In the future, we are interested in also performing the genetic algorithm step of our algorithm for power minimization which would help us better learn the low-power regions. We are also interested in reducing the root mean squared error further by exploring more sophisticated prediction methods like neural networks. Furthermore, we are interested in extending our algorithm for learning the gradient of the statistical power manifold for different parameters.

Acknowledgments This research was supported by the National Institutes of Health research grant R01 AA027264 (PIs: Lane/Hennes).

References

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2021). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods, 26*(3), 295.

Bakker, M., Hartgerink, C. H., Wicherts, J. M., & van der Maas, H. L. (2016). Researchers' intuitions about power in psychological research. *Psychological Science, 27*(8), 1069–1077.

Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543–554.

Cafri, G., Kromrey, J. D., and Brannick, M. T. (2010). A meta-meta-analysis: Empirical review of statistical power, type i error rates, effect sizes, and model selection of meta-analyses published in psychology. *Multivariate Behavioral Research, 45*(2), 239–270.

Cohen, J. (1992). Things i have learned (so far). In *Annual Convention of the American Psychological Association, 98th, Aug, 1990, Boston, MA, US; Presented at the Aforementioned Conference.* American Psychological Association.

Fraley, R. C., & Vazire, S. (2014). The n-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS One, 9*(10), e109019.

Goldberg, D. E., & Holland, J. H. (1988). Genetic algorithms and machine learning. *Machine Learning, 3*, 95–99.

Holland, J. H. (1992). Genetic algorithms. *Scientific American, 267*(1), 66–73.

Lane, S. P., & Hennes, E. P. (2018). Power struggles: Estimating sample size for multilevel relationships research. *Journal of Social and Personal Relationships, 35*(1), 7–31.

Lane, S. P., & Hennes, E. P. (2019). Conducting sensitivity analyses to identify and buffer power vulnerabilities in studies examining substance use over time. *Addictive Behaviors, 94*, 117–123.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods, 9*(2), 147.

Mitchell, M. (1998). *An introduction to genetic algorithms*. MIT Press.

Rast, P., & Hofer, S. M. (2014). Longitudinal design considerations to optimize power to detect variances and covariances among rates of change: simulation results based on actual longitudinal studies. *Psychological Methods, 19*(1), 133.

Umrawal, A. K., Lane, S. P., & Hennes, E. P. (2022). A genetic algorithm-based framework for learning statistical power manifold. *Preprint. arXiv:2209.00215*.

Yang, Y., Hillebrand, H., Lagisz, M., Cleasby, I., & Nakagawa, S. (2022). Low statistical power and overestimated anthropogenic impacts, exacerbated by publication bias, dominate field studies in global change biology. *Global Change Biology, 28*(3), 969–989.

Using Nonparametric Mixture Models to Model Effect Heterogeneity in Meta-analysis of Very Rare Events

Heinz Holling and Katrin Jansen

Abstract Modeling heterogeneity in meta-analysis of count data is challenging when the event of interest is rare. Then, models based on the assumption of a normal random-effects distribution often fail to detect heterogeneity or to provide unbiased estimates of the between-study variance. The aim of this study is to evaluate the performance of logistic and log-linear nonparametric mixture models in detecting heterogeneity, estimating the pooled effect size, and estimating the between-study variance in meta-analysis of very rare events. These models do not require a parametric specification of the random-effects distribution. Performance was evaluated by means of a simulation study in which the number of primary studies, the sample size within studies, the mixture component weights and the baseline probabilities as well as the effect sizes of the components were varied. The results show that nonparametric mixture models perform well in terms of parameter estimation as long as enough studies with large sample sizes are available. Large numbers of studies and large sample sizes are required for a reliable detection of heterogeneity, in particular when non-zero effects are associated with a low occurrence probability.

Keywords Meta-analysis · Nonparametric mixture model · Rare events · Count data analysis · Heterogeneity

1 Introduction

A meta-analysis is a quantitative summary of studies on the same research question. It is typically conducted in form of a synthesis of quantitative measures of an effect of interest, so-called effect sizes, which are extracted from the individual studies. Here, we will focus on meta-analysis of count data. Specifically, we focus on a

H. Holling (✉) · K. Jansen
University of Münster, Department of Psychology, Münster, Germany
e-mail: holling@uni-muenster.de; katrinjansen@uni-muenster.de

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_18

Table 1 Contingency table

	Event	No event
Treatment	y_{i1}	$n_{i1} - y_{i1}$
Control	y_{i0}	$n_{i0} - y_{i0}$

situation in which for each study i, $i = 1, ..., k$, the occurrence of an event of interest was assessed for n_{i1} subjects in a treatment group and n_{i0} subjects in a control group. The observations from study i can be summarized in a contingency table (see Table 1 for illustration).

From these data, an effect size can be computed, such as the log odds ratio (log OR),

$$\log(\widehat{OR}_i) = \log\left(\frac{y_{i1}/(n_{i1} - y_{i1})}{y_{i0}/(n_{i0} - y_{i0})}\right) \tag{1}$$

or the log relative risk (log RR),

$$\log(\widehat{RR}_i) = \log\left(\frac{y_{i1}/n_{i1}}{y_{i0}/n_{i0}}\right) \tag{2}$$

In the following, study effect sizes will be denoted by $\hat{\theta}_i$, regardless of whether the log OR or the log RR is used. In a conventional meta-analysis using the inverse variance model, it is typically assumed that $\hat{\theta}_i$ follows a normal distribution with mean θ_i and variance σ_i^2. For the log OR, σ_i^2 can be estimated by

$$\widehat{\text{Var}}(\hat{\theta}_i) = 1/y_{i1} + 1/(n_{i1} - y_{i1}) + 1/y_{i0} + 1/(n_{i0} - y_{i0}) \tag{3}$$

and for the log RR by

$$\widehat{\text{Var}}(\hat{\theta}_i) = 1/y_{i1} - 1/n_{i1} + 1/y_{i0} - 1/n_{i0}. \tag{4}$$

In a meta-analysis, we are interested in obtaining a quantitative summary of these study effect sizes in form of a pooled effect size and a confidence interval. A further objective is to model the heterogeneity among the true effect sizes θ_i which are usually assumed to follow a normal distribution with mean θ and variance τ^2, such that we obtain the conventional random-effects model for meta-analysis. The maximum likelihood estimator for the pooled effect obtained from this model is $\hat{\theta} = \sum_{i=1}^{k} w_i\hat{\theta}_i / \sum_{i=1}^{k} w_i$, where $w_i = 1/(\sigma^2 + \tau^2)$. Several estimators for τ^2, e.g., the restricted maximum likelihood estimator have been proposed (see Borenstein et al., 2009, for a general introduction to meta-analysis).

Problems with the conventional random-effects model occur when any of the four cells in the contingency table is zero: Then, study effect sizes as well as their variances are no longer defined. Even if an ad-hoc fix to this problem is used by applying continuity corrections (see Sweeting et al., 2004, for a discussion), the

assumption of a normal distribution within studies is questionable when event probabilities are small (Jackson & White, 2018). Finally, estimation of heterogeneity is challenging in meta-analysis of rare events (Zhang et al., 2020).

Commonly used alternatives to the conventional random effects model in rare events meta-analysis are the Mantel-Haenszel method (Mantel & Haenszel, 1959) and the Peto method (Yusuf et al., 1985), both of which are based on the assumption of fixed effects (i.e., $\theta_1 = ... = \theta_k$). However, simulation studies have shown that these models yield biased estimates and unsatisfactory coverage of confidence intervals when study effect sizes are truly heterogeneous (Bhaumik et al., 2012; Kuss, 2014).

Generalized linear mixed models (GLMMs) are important alternative random effects models for count data. These models are typically based on the assumption that the counts of the individual studies follow binomial distributions or Poisson distributions (see Beisemann et al. (2020) for an overview of models for the log RR, and Jansen and Holling (2022), for an overview of models for the log OR). GLMMs incorporate the random effects by assuming a normal distribution for the effect size parameter. These distributional assumptions for the random effects might not hold and can almost never be tested in the context of meta-analysis.

In this chapter, we propose nonparametric mixture models as a potential solution to this problem, which allow for modeling heterogeneity without requiring the assumption of a parametric random-effects distribution. By using a finite mixture model we leave the distribution of the true effect sizes unspecified. In a previous simulation study (Holling et al., 2022), these nonparametric mixture models were investigated in the context of rare events meta-analysis for count data and it was shown that they perform well when certain requirements with regard to sample sizes and numbers of studies are met. Here, we evaluate whether these models also prove to be useful for meta-analysis of very rare events, i.e., in an even more challenging situation.

The remainder of this chapter is structured as follows: In the following section, we describe the logistic and log-linear nonparametric mixture model. Afterwards, we present the design and the results of a simulation study evaluating these models in the context of meta-analysis of very rare events. Finally, we conclude the chapter with a short discussion.

2 Logistic and Log-Linear Nonparametric Mixture Models

In the following, we assume that we have collected count data from k studies which have observed the number of events in a treatment group and a control group, such that the data can be summarized in a contingency table (see Table 1, for an example).

For the logistic mixture model, we assume that the observed counts are draws from random variables Y_{ij} which follow binomial distributions $\text{Bin}(n_{ij}, p_{ij})$. We set up the following model equation:

$$\log\left(\frac{E(Y_{ij})}{n_{ij} - E(Y_{ij})}\right) = \alpha_i + \beta_i \times j. \tag{5}$$

Note that the slope, β_i, represents the log OR of study i.

For the log-linear mixture model, we assume that the observed counts are draws from random variables Y_{ij} which follow Poisson distributions Poi(λ_{ij}). We can then set up the model equation:

$$\log\left(E(Y_{ij})\right) = \alpha_i + \beta_i \times j + \log(n_{ij}). \tag{6}$$

Here, the slope β_i represents the log RR of study i.

If we were to specify a GLMM, we would now make the assumption that α_i and β_i follow normal distributions. As outlined above, we do not want to assume a parametric random-effects distribution here, and so, we leave the distribution of (α_i, β_i) unspecified and simply assume that they follow some mixing distribution Q.

We obtain the following log-likelihood for the logistic mixture model:

$$l(Q) = \sum_i \log\left[\int \prod_j p(y_{ij}; n_{ij}, \text{expit}(\alpha_i + \beta_i \times j))Q(d\alpha_i, d\beta_i)\right], \tag{7}$$

where $p(\cdot)$ is the probability density of a binomial distribution and $\text{expit}(x) = \exp(x)/\{1+\exp(x)\}$. Lindsay (1983, 1995) has shown that the maximum likelihood estimator obtained from Eq. (7) is always discrete, such that, without loss of generality, Eq. (7) can be replaced by

$$l(Q) = \sum_i \log\left[\sum_{s=1}^{S} \prod_j p(y_{ij}; n_{ij}, \text{expit}(\alpha_s + \beta_s \times j))q_s\right], \tag{8}$$

where $q_1, ..., q_s$ are referred to as the weights of the discrete mixture log-likelihood. Equations (7) and (8) refer to the log-likelihood of a model with heterogeneous effect sizes, for which $\beta_s \neq \beta_{s'}$ for $s \neq s'$. Making the restriction $\beta_s = \beta$ for $s = 1, ..., S$ yields the likelihood of the corresponding model with homogeneous effects, in which only the intercepts α_s vary across components. In close analogy, we obtain the following log-likelihood for a log-linear mixture model:

$$l(Q) = \sum_i \log\left[\sum_{s=1}^{S} \prod_j p(y_{ij}; \exp(\alpha_s + \beta_s \times j + \log(n_{ij})))q_s\right]. \tag{9}$$

For a given number of components S, the log-likelihoods in Eqs. (8) and (9) can be easily maximized using the EM algorithm (Dempster et al., 1977; McLachlan,

2008). In typical applications of the nonparametric mixture model, S is unknown. Different solutions to this problem have been proposed: One option is to start with a model where $S = 1$, and then sequentially increase the number of components until no further increase in the log-likelihood is detected. Another option is to use fit indices, such as the Akaike information criterion (AIC), or the Bayesian information criterion (BIC), to obtain the best fitting model from a predefined set of models with different numbers of components.

Apart from the parameter estimates for the single components, $\hat{q}_s$, $\hat{\alpha}_s$ and $\hat{\beta}_s$, $s = 1, ..., S$, we can also obtain estimates for the pooled effect:

$$\hat{\bar{\beta}} = \sum_{s=1}^{S} \hat{q}_s \hat{\beta}_s, \tag{10}$$

and the between-study variance:

$$\hat{\tau}^2 = \sum_{s=1}^{S} \hat{q}_s (\hat{\beta}_s - \hat{\bar{\beta}})^2. \tag{11}$$

Both logistic and log-linear nonparametric mixture models were investigated in simulation studies by Holling et al. (2022) in the context of rare events meta-analysis. These studies showed that model selection using the BIC and parameter estimation works well for studies with large sample sizes (such as $n = 1000$). For small sample sizes (such as $n = 100$), the AIC performed better with regard to model selection. Performance in terms of parameter estimation was reasonably good for small samples in case of a large number of studies ($k \geq 40$) or when components were more distinct in terms of their baseline probabilities or in terms of their component effect sizes. However, Holling et al. (2022) only investigated performance for baseline event probabilities of 0.05 and above. Simulation studies have shown that estimation in parametric random effects models is most challenging for events that occur even less often, at event probabilities of 0.01 and below (Jackson et al., 2018). In this chapter, we investigate the performance of nonparametric mixture models for meta-analyses in which the baseline event probability is 0.01 or smaller. A detailed introduction to nonparametric mixture models can be found in Böhning (2000).

3 Simulation Study

The simulation study was conducted in `R` (R Core Team, 2021) and run on the computing cluster PALMA II (https://www.uni-muenster.de/ZIV/Technik/Server/HPC.html) at the University of Münster. Computations were parallelised using the `doParallel` package (Microsoft Corporation and Weston, 2022). Separate simulation studies were conducted to evaluate logistic and log-linear mixture

models. The two simulation studies differed only with regard to how the effect size was defined, but were otherwise conducted using the same parameter values and data generating mechanism, as described below. The code and data from this simulation study are available at https://osf.io/u2bda/.

For all simulation conditions, the true number of components was $S = 2$. We varied the number of studies k, the total sample size of each study n, and the probability of component 1 q_1, using the values given in Table 2. Meta-analyses of rare events are often based on as few as ten studies (Davey et al., 2011). Since it is trivial that nonparametric mixture models will not perform well if one component consists of (almost) only double-zero studies, it is desirable to avoid that such data are generated in many simulation replications. However, with as few as ten studies, such a scenario is likely to occur. Therefore, we chose to simulate conditions with more than ten studies, specifically, $k = 30$ and $k = 60$. In our previous simulation study on rare events (Holling et al., 2022), nonparametric mixture models performed well in terms of model selection and parameter estimation for $n = 1000$, while for sample sizes as small as $n = 100$, model selection performance was impaired. For $n = 100$, we would expect to see similar or even worse performance drawbacks when events are very rare. We thus decided to include conditions with sample sizes as large as $n = 1000$ as well as conditions with sample sizes in between $n = 100$ and $n = 1000$ (specifically, we chose $n = 250$ and $n = 500$). The component mixture weights q_s were chosen to mirror a scenario with balanced components and ($q_1 = 0.5$) and a scenario with unbalanced components ($q_1 = 0.3$). The values for the component baseline probabilities $p_{0,1}$ and $p_{0,2}$ were 0.005 and 0.01, where $p_{0,s}$ is the baseline probability of component s. These values were chosen to model a very rare event, in line with the classification by Jackson et al. (2018). The component effect sizes β_1 and β_2 were $\log(1) = 0$ (corresponding to a neutral effect size) and $\log(3) \approx 1.10$, reflecting a moderate degree of heterogeneity. Simulation conditions were defined by fully crossing the parameters given in Table 2. Note that for conditions in which $q_1 = 0.5$, the component labels s of β_s were not permuted since this would generate a set of equivalent conditions resulting from the permutation of the component labels of $p_{0,s}$ and the fact that $q_1 = q_2$. This approach resulted in a total number of 36 simulation conditions.

To mirror randomization, the sample size of the treatment group, n_{i1} was drawn from a binomial distribution with sample size n and event probability 0.5. The sample size of the control group was then determined as $n_{i0} = n - n_{i1}$. Each

Table 2 Simulation conditions

Parameter	Values
Number of studies k	30, 60
Sample sizes n	250, 500, 1000
Component 1 probability q_1	0.3, 0.5
Baseline probabilities $p_{0,s}$	$p_{0,1} = 0.005$ & $p_{0,2} = 0.01$, $p_{0,1} = 0.01$ & $p_{0,2} = 0.005$
Effect sizes β_s	$q_1 = 0.5$: $\beta_1 = \log(1)$ & $\beta_2 = \log(3)$
	$q_1 = 0.3$: $\beta_1 = \log(1)$ & $\beta_2 = \log(3)$, $\beta_1 = \log(3)$ & $\beta_2 = \log(1)$

study was assigned to one of the two components based on a random draw from a Bernoulli distribution with probability q_1. The component treatment group event probabilities $p_{1,s}$ were calculated from β_s and $p_{0,s}$. Observations for each group and each study were drawn from binomial distributions with sample sizes n_{ij}, $i = 1, ..., k$, $j \in \{0, 1\}$ and event probabilities $p_{j,s}$, where s was the component the study had been assigned to.

Logistic and log-linear nonparametric mixture models, as defined by Eqs. (8) and (9) in Sect. 2, were fitted using the `flexmix` package (Grün and Leisch, 2008). Within `flexmix`, we used the function `stepFlexmix`, which fits the model repeatedly for different values of S and returns the maximum likelihood solution for each value of S. We fitted models with $S = 1$, $S = 2$ and $S = 3$. In addition to models with heterogeneous effects, we fitted models with homogeneous effects for which the restriction $\beta_s = \beta$ is made, as described in Sect. 2. This results in five models to be considered in model selection (since the model with $S = 1$ assumes homogeneous effects per definition). Considering models with homogeneous effects in model selection entails that if the correct model is identified, this means that heterogeneity has been correctly detected. For each model, the number of repetitions was set to 10. In each repetition, starting values were defined by random assignment of the observations to the components.

Performance was evaluated separately for logistic and log-linear mixture models in terms of (1) the accuracy of model selection by the AIC and BIC, and (2) parameter estimation, assessed by (a) mean bias, (b) median bias and (c) the standard deviation of $\hat{\bar{\beta}}$ and $\hat{\tau}^2$, respectively. The mean bias of $\hat{\bar{\beta}}$ and $\hat{\tau}^2$ is defined as

$$\text{Mean bias}_{\hat{\bar{\beta}}} = \sum_{i=1}^{R} \frac{\hat{\bar{\beta}} - \bar{\beta}}{R} \tag{12}$$

and

$$\text{Mean bias}_{\hat{\tau}^2} = \sum_{i=1}^{R} \frac{\hat{\tau}^2 - \tau^2}{R}, \tag{13}$$

respectively, where R is the number of simulation replications. The median bias of $\hat{\bar{\beta}}$ and $\hat{\tau}^2$ results as the median of $\left(\hat{\bar{\beta}}_1 - \bar{\beta}, ..., \hat{\bar{\beta}}_R - \bar{\beta}\right)$ and the median of $\left(\hat{\tau}_1^2 - \tau^2, ..., \hat{\tau}_R^2 - \tau^2\right)$, respectively.

4 Results

In Table 3, we present the results in terms of model selection for the logistic model separately for simulation conditions in which a small component baseline probability was associated with a large component effect size (i.e., $p_{0,s} = 0.005$

Table 3 Model selection performance (logistic mixture model)

k	n	$p_{0,s} = 0.005$ and $\beta_s = \log(3)$	AIC	BIC
30	250	Yes	0.08–0.10	0.01–0.01
30	500	Yes	0.17–0.22	0.03–0.04
30	1000	Yes	0.40–0.49	0.11–0.17
60	250	Yes	0.11–0.15	0.01–0.01
60	500	Yes	0.28–0.35	0.03–0.04
60	1000	Yes	0.62–0.75	0.20–0.33
30	250	No	0.36–0.55	0.14–0.31
30	500	No	0.66–0.82	0.49–0.73
30	1000	No	0.85–0.89	0.82–0.95
60	250	No	0.55–0.77	0.26–0.53
60	500	No	0.83–0.90	0.72–0.93
60	1000	No	0.87–0.89	0.95–0.97

Table 4 Model selection performance (log-linear mixture model)

k	n	$p_{0,s} = 0.005$ and $\beta_s = \log(3)$	AIC	BIC
30	250	Yes	0.08–0.10	0.01–0.01
30	500	Yes	0.17–0.22	0.02–0.03
30	1000	Yes	0.38–0.49	0.10–0.16
60	250	Yes	0.11–0.13	0.01–0.01
60	500	Yes	0.26–0.34	0.02–0.05
60	1000	Yes	0.64–0.75	0.21–0.32
30	250	No	0.35–0.55	0.13–0.30
30	500	No	0.67–0.83	0.50–0.72
30	1000	No	0.87–0.89	0.83–0.93
60	250	No	0.54–0.78	0.26–0.54
60	500	No	0.85–0.90	0.74–0.93
60	1000	No	0.88–0.90	0.94–0.96

and $\beta_s = \log(3)$ for $s = 1$ or $s = 2$) and simulation conditions in which a small component baseline probability was associated with a neutral component effect size (i.e., $p_{0,s} = 0.005$ and $\beta_s = \log(1)$ for $s = 1$ or $s = 2$). The corresponding results for the log-linear model are presented in Table 4. Results for logistic and log-linear models are similar and will thus be described simultaneously.

For the conditions in which $p_{0,s} = 0.005$ and $\beta_s = \log(3)$ for $s = 1$ or $s = 2$, (presented in the upper half of Tables 3 and 4), both the AIC and the BIC perform poor in terms of model selection. Only if k and n are very large, the AIC achieves a somewhat satisfactory performance and consistently selects the correctly specified model in more than 60% of simulation replications. In simulation conditions with $p_{0,s} = 0.005$ and $\beta_s = \log(1)$ for $s = 1$ or $s = 2$ (presented in the lower half of Tables 3 and 4), model selection performance is notably better and can be considered

satisfactory if either k or n is large. Note that in most conditions, the AIC performs better in model selection than the BIC.

Next, we evaluate the estimation of $\bar{\beta}$ for the model which was correctly specified (i.e., a model with heterogeneous effects and $S = 2$). Results in terms of mean bias, median bias and $SD(\hat{\bar{\beta}})$ are presented in Table 5 for the logistic model, grouped by the number of studies k and the sample size n. Analogous results for the log-linear model are shown in Table 6. The results for both models are similar and will be described simultaneously. Note that we removed simulation replications from the analysis in which β_1 or β_2 exceeded $\log(1000)$, since estimates of this magnitude can point to issues in maximum likelihood estimation in the analysis of sparse data (see Heinze, 2006). The range of the number of simulation replications which the results are based on is given in column 3 of Table 5, for the logistic mixture model, and in column 3 of Table 6, for the log-linear mixture model. We see that both mean and median bias of $\hat{\bar{\beta}}$ are generally small, in particular when k and n are large. The larger k and n, the more efficient becomes the estimation of $\hat{\bar{\beta}}$, as can be assessed by evaluating $SD(\hat{\bar{\beta}})$.

Finally, we evaluate how well the between-study variance, τ^2, is estimated by the correctly specified model. Outliers were removed from the analysis in the same way as in the analysis of the estimation of $\bar{\beta}$. Results in terms of mean bias, median bias and $SD(\hat{\tau}^2)$ are displayed in Table 7, for the logistic mixture model, and in Table 8, for the log-linear mixture model. Since the results of the two models are similar, we will describe them simultaneously. While median bias is mostly adequate even for small values of k and n, small sample sizes are associated with a large mean bias and inefficient estimation of τ^2. Even for $k = 60$ and $n = 500$, the standard deviation of

Table 5 Estimation of $\bar{\beta}$ (logistic mixture model)

k	n	Replications	Mean bias	Median bias	SD
30	250	3733–4739	−0.022–0.045	−0.002–0.015	0.359–0.440
30	500	4279–4997	0.000–0.024	−0.008–0.010	0.208–0.256
30	1000	4718–5000	0.001–0.010	−0.004–0.007	0.153–0.172
60	250	4265–4971	−0.008–0.046	−0.008–0.016	0.257–0.340
60	500	4585–5000	−0.001–0.021	−0.001–0.011	0.147–0.188
60	1000	4877–5000	0.002–0.005	0.000–0.007	0.108–0.124

Table 6 Estimation of $\bar{\beta}$ (log-linear mixture model)

k	n	Replications	Mean bias	Median bias	SD
30	250	3745–4748	−0.022–0.051	−0.003–0.023	0.371–0.441
30	500	4268–4998	0.000–0.027	−0.006–0.010	0.211–0.260
30	1000	4741–5000	0.003–0.012	0.005–0.007	0.153–0.178
60	250	4287–4987	−0.012–0.041	−0.002–0.008	0.253–0.348
60	500	4559–5000	0.003–0.019	0.000–0.011	0.144–0.183
60	1000	4883–5000	0.000–0.005	0.001–0.003	0.108–0.122

Table 7 Estimation of τ^2 (logistic mixture model)

k	n	Replications	Mean bias	Median bias	SD
30	250	3733–4739	0.327–0.608	0.006–0.045	1.173–1.559
30	500	4279–4997	0.045–0.227	−0.001–0.072	0.243–0.684
30	1000	4718–5000	0.011–0.057	−0.010–0.026	0.136–0.264
60	250	4265–4971	0.124–0.521	0.009–0.070	0.633–1.389
60	500	4585–5000	0.022–0.162	−0.004–0.058	0.142–0.565
60	1000	4877–5000	0.003–0.032	−0.007–0.021	0.093–0.160

Table 8 Estimation of τ^2 (log-linear mixture model)

k	n	Replications	Mean bias	Median bias	SD
30	250	3745–4748	0.348–0.629	−0.006–0.040	1.223–1.665
30	500	4268–4998	0.048–0.215	−0.004–0.060	0.235–0.668
30	1000	4741–5000	0.009–0.059	−0.010–0.034	0.133–0.240
60	250	4287–4987	0.117–0.503	0.003–0.031	0.545–1.395
60	500	4559–5000	0.019–0.149	−0.007–0.049	0.141–0.518
60	1000	4883–5000	0.005–0.027	−0.004–0.013	0.093–0.164

$\hat{\tau}^2$ can still be as large as 0.5 depending on the combination of $p_{0,s}$ and β_s. Thus, a proper estimation of τ^2 seems to require considerably larger numbers of studies and sample sizes than a proper estimation of $\hat{\hat{\beta}}$.

5 Conclusion and Discussion

Nonparametric mixture models have been successfully applied to statistical issues in many disciplines, e.g., medicine, psychology or economics. Applications to different problems in meta-analysis are provided by, e.g., Doebler and Holling (2014), Holling et al. (2012) and Malzahn et al. (2012). Holling et al. (2022) developed nonparametric mixture models for meta-analysis of count data based on log-linear and logistic regression. Simulation studies including rare data provided good estimates of both the pooled effect sizes and heterogeneity when the assumptions of the models were fulfilled and enough studies with reasonably large sample sizes were available. In this chapter, we explored the potential of these nonparametric mixture models for meta-analysis of very rare events. By using a binomial or Poisson model within studies, these models are adequate for count data even when some studies have zero counts in either or both study groups, a situation which is frequently encountered in meta-analysis of very rare events and is problematic for conventional meta-analysis models. On top of that, nonparametric mixture models provide a flexible way to account for heterogeneity in meta-analysis of count data. They avoid the assumption of a normal random-effects distribution—

an assumption which usually cannot be tested and may be questionable, for instance when the model is not correctly specified.

We found that even for events with occurrence probabilities below 0.01, a sensible model selection performance can be achieved in some situations. Problems with regard to selection of the correct model occur when non-zero effect sizes are associated with small baseline probabilities. When baseline probabilities are small, it is more likely that double zero studies occur. When double zero studies occur in a component with a neutral effect size, this is less problematic as double zero studies indicate that there is no treatment effect. However, when double zero studies occur in a component with a non-zero effect size, they might induce bias towards a neutral effect. In our simulation study, there was always one component with a neutral effect and one component with a non-zero effect. If double zero studies in the latter component induce bias towards a neutral effect, this might result in a better fit for a model with one component, and as a consequence, a model with one component is selected by model selection criteria. In summary, it is likely that the poor model selection performance in conditions in which a small baseline probability is paired with a non-zero effect is caused by many double zero studies in the respective component. For larger sample sizes, we would expect to see a better model selection performance even under such challenging conditions, since larger sample sizes make the occurrence of double zero studies less likely. We found that in meta-analysis of very rare events, the AIC outperforms the BIC in terms of model selection. However, since the BIC is asymptotically consistent in model selection while the AIC is not, it seems likely that this result would not generalize to larger numbers of studies and larger sample sizes (see Vrieze, 2012, for a discussion of the differences between the AIC and BIC).

In terms of parameter estimation, nonparametric mixture models perform well also for very rare events as long as the number of studies and the sample sizes are large enough. As expected, more studies and larger samples are required for meta-analysis of very rare events as compared to meta-analysis of rare events (cf. Holling et al., 2022). In addition, it is important to note that a proper estimation of the heterogeneity variance requires considerably larger numbers of studies and sample sizes than a proper estimation of the pooled effect size.

This study was the first evaluation of nonparametric mixture models for meta-analysis of count data with very rare events. Future studies should further investigate the performance of these models. In particular, the conditions of our simulation study could be extended. Interesting extensions would include, for instance, the investigation of further combinations of component baseline probabilities and component effect sizes. Furthermore, the ability of correctly clustering the studies could serve as another criterion of performance as nonparametric mixture models also provide the means to assign observations (here studies) to clusters. Moreover, it would be interesting to re-analyze existing meta-analyses including rare and very rare events, e.g., from the Cochrane Library, by using nonparametric mixture modeling to explore whether this approach might be statistically more appropriate or leads to more plausible interpretations of the results.

So far, nonparametric mixture models have not received much attention in the context of meta-analysis, although they have the advantage of not requiring parametric assumptions with regard to the random-effects distribution. This flexibility could prove useful in situations in which the common intercept-only model, as used in conventional random effects meta-analysis using the inverse variance model, is overly simplistic, for instance in the presence of moderation effects which cannot be modeled explicitly due to missing data. Hence, practitioners should explore whether this approach might be more appropriate for the analysis of their meta-analytic count data, especially for rare and very rare events.

References

Beisemann, M., Doebler, P., & Holling, H. (2020). Comparison of random-effects meta-analysis models for the relative risk in the case of rare events: A simulation study. *Biometrical Journal, 62*(7), 1597–1630.

Bhaumik, D. K., Amatya, A., Normand, S.-L. T., Greenhouse, J., Kaizar, E., Neelon, B., & Gibbons, R. D. (2012). Meta-analysis of rare binary adverse event data. *Journal of the American Statistical Association, 107*(498), 555–567.

Böhning, D. (2000). *Computer-assisted analysis of mixtures and applications: meta-analysis, disease mapping, and others*. Chapman & Hall/CRC, Boca Raton.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: John Wiley & Sons.

Davey, J., Turner, R. M., Clarke, M. J., & Higgins, J. P. (2011). Characteristics of meta-analyses and their component studies in the cochrane database of systematic reviews: a cross-sectional, descriptive analysis. *BMC Medical Research Methodology, 11*(1), 1.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–22.

Doebler, P., & Holling, H. (2014). Meta-analysis of diagnostic accuracy and ROC curves with covariate adjusted semiparametric mixtures. *Psychometrika, 80*(4), 1084–1104.

Grün, B., & Leisch, F. (2008). FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software, 28*(4), 1–35.

Heinze, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in Medicine, 25*(24), 4216–4226.

Holling, H., Böhning, W., & Böhning, D. (2012). Likelihood based clustering of meta-analytic SROC curves. *Psychometrika, 77*(1), 106–126.

Holling, H., Jansen, K., Böhning, W., Böhning, D., Martin, S., & Sangnawakij, P. (2022). Estimation of effect heterogeneity in rare events meta-analysis. *Psychometrika, 87*(3), 1081–1102.

Jackson, D., Law, M., Stijnen, T., Viechtbauer, W., & White, I. R. (2018). A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Statistics in Medicine, 37*(7), 1059–1085.

Jackson, D., & White, I. R. (2018). When should meta-analysis avoid making hidden normality assumptions? *Biometrical Journal, 60*(6), 1040–1058.

Jansen, K., & Holling, H. (2022). Random-effects meta-analysis models for the odds ratio in the case of rare events under different data-generating models: A simulation study. *Biometrical Journal, 65*(3), e2200132.

Kuss, O. (2014). Statistical methods for meta-analyses including information from studies without any events-add nothing to nothing and succeed nevertheless. *Statistics in Medicine, 34*(7), 1097–1116.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: A general theory. *The Annals of Statistics, 11*(1), 86–94.

Lindsay, B. G. (1995). Mixture models: Theory, geometry and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics, 5*, i–163.

Malzahn, U., Böhning, D., & Holling, H. (2012). Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika, 87*(3), 619–632.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *JNCI: Journal of the National Cancer Institute, 22*(4), 719.

McLachlan, K. (2008). *The EM algorithm and its extensions*. New Jersey: John Wiley & Sons.

Microsoft Corporation, & Weston, S. (2022). *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. R package version 1.0.17.

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Sweeting, M. J., Sutton, A. J., & Lambert, P. C. (2004). What to add to nothing? use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine, 23*(9), 1351–1375.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17*(2), 228–243.

Yusuf, S., Peto, R., Lewis, J., Collins, R., & Sleight, P. (1985). Beta blockade during and after myocardial infarction: An overview of the randomized trials. *Progress in Cardiovascular Diseases, 27*(5), 335–371.

Zhang, C., Chen, M., & Wang, X. (2020). Statistical methods for quantifying between-study heterogeneity in meta-analysis with focus on rare binary events. *Statistics and Its Interface, 13*(4), 449–464.

Investigating Differential Item Functioning via Odds Ratio in Cognitive Diagnosis Models

Ya-Hui Su and Tzu-Ying Chen

Abstract The increasing number of tests being developed has prompted more people to investigate the association between test items and skill attributes and state of knowledge, spurring the development of the cognitive diagnosis models. Several studies have predominantly adopted the Mantel–Haenszel (MH) method to detect differential item functioning (DIF) under such models. Jin et al. (2018) used the odds ratio (OR) method to examine DIF under the Rasch model, which assumed latent traits were continuous. It was found that that the OR method outperformed the traditional MH method in terms of type I error rate control and statistical power. However, no studies have applied the OR method in DIF detection under the cognitive diagnosis models. Therefore, this study investigated the effectiveness of DIF detection methods, including the MH method, MH method with purification procedure, MH method with attribute patterns as the matching variables, OR method, and OR method with purification procedure. According to the results, the effectiveness of DIF detection was affected by sample size and the proportion of DIF items; specifically, a large sample size and a high proportion of DIF items were associated with increased and decreased statistical power, respectively. The purification procedure enhanced the DIF detection effectiveness and reduced the type I error rate in both the OR and MH methods.

Keywords Cognitive diagnostic model · Differential item functioning · MH · Odds ratio

1 Introduction

Cognitive diagnosis models (CDMs) can be used to obtain diagnostic information, which specifies whether the required skills has been mastered (Hartz, 2002; Junker

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

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_19

& Sijtsma, 2001; Mislevy et al., 2000; Rupp et al., 2010; Tatsuoka, 1983). The CDMs have been applied mainly in the educational context to provide teachers and students with information concerning if each of a group of specific skills has been mastered. The skills here are often referred to as attributes. To make sure the quality of a test, the detection of differential item functioning (DIF) is a critical procedure for examining test fairness. Many DIF detection methods have been used in CDMs (Hou et al., 2014; Sünbül, 2019; Zhang, 2006), including Mantel-Haenszel (MH; Holland & Thayer, 1986, 1988; Mantel & Haenszel, 1959), simultaneous item bias test (SIBTEST; Shealy & Stout, 1993), logistic regression (LR; Swaminathan & Rogers, 1990), and Wald test (Morrison, 1967), etc. Among these DIF methods, the MH method is predominantly used in CDMs.

Jin et al. (2018) used the odds ratio (OR) method to examine DIF under the Rasch model, which assumed that latent traits were continuous. It was found that the OR method outperformed the MH method in terms of type I error rate control and statistical power. However, no studies have applied the OR method for DIF detection under the CDMs. Therefore, this study investigated the effectiveness of DIF detection obtained by the MH method, MH method with purification procedure, MH method with attribute patterns as the matching variables, OR method, and OR method with purification procedure in CDMs.

2 Method

In the study, simulations were conducted via RStudio 4.0.0 and difR package. Please contact the authors for accessing code if you are interested. Similar to de la Torre (2011), the Q matrix with five attributes were considered and generated in the study. Previous CDM studies used different methods to generate examinees' attribute patterns (Chiu & Douglas, 2013; de la Torre & Douglas, 2004; Hou et al., 2014; Li, 2008; Li & Wang, 2015; Zhang, 2006). In this study, examinees' attribute patterns were generated according to Zhang (2006). An examinee had five attributes, and decided each attribute of the attribute patterns was mastered if the z scores randomly drawn from the multivariate normal distribution were larger than zero.

Several CDMs have been proposed in the literature (Haertel, 1989; Junker & Sijtsma, 2001; Maris, 1999; Roussos et al., 2007; Templin & Henson, 2006). The present study focused only on the deterministic input, noisy, and gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model, and the deterministic input, noisy, or gate (DINO; Templin & Henson, 2006) model. The DINA model assumes that each attribute measured by an item must be successfully applied in order to answer an item correctly. The DINA model includes two item parameters, s_j and g_j. s_j is the slip parameter, which represents the probability that an examinee who has all the required attributes fails to answer item j correctly; and g_j is the guessing parameter, which represents the probability that an examinee who lacks at least one of the required attributes answers item j correctly. The probability of answering an item correctly can be written as

$$P\left(X_{ij}=1|s_j,g_j,\eta_{ij}\right)=\left(1-s_j\right)^{\eta_{ij}}g_j{}^{(1-\eta_{ij})}, \tag{1}$$

where $\eta_{ij}=\prod_{k=1}^{K}\alpha_{ik}{}^{q_{jk}}$ represents that examinee i has mastered all of the required attributes for item j, and K is the total number of k attributes. In contrast to the DINA model, the DINO model is a compensatory CDM. The DINO model also includes two item parameters, s_j and g_j.The probability of answering an item correctly can be written as

$$P\left(X_{ij}=1|s_j,g_j,\eta_{ij}\right)=\left(1-s_j\right)^{\omega_{ij}}g_j{}^{(1-\omega_{ij})}, \tag{2}$$

where $\omega_{ij}=1-\prod_{k=1}^{K}(1-\alpha_{ik})^{q_{jk}}$ represents that examinee i has mastered at least one measured attribute for item j. In the DINO model, s_j represents the probability of not answering item j correctly when an examinee has mastered at least one measured attribute; g_j represents the probability of answering item j correctly when an examinee has not mastered all measured attributes. In this study, data were generated according to the DINA and the DINO models.

Zhang (2006) investigated the effect of sample size on the DIF detection in CDM. In practice, different sample size for the reference and focal groups is very common. Thus, the sample size for reference and focal groups were 500, 1000, and 2000. In this study, six different levels of sample size were manipulated. Previous studies have been investigated the effect of the item parameters on DIF detection (Hou et al., 2014; Li, 2008). To investigate the effect of quality of item pool on DIF detection in CDM. For the reference group, the item parameters s_j and g_j were set to be $s_j=g_j=.25$ in the high-quality item bank, and $s_j=g_j=.75$ in the low-quality item bank. Many DIF studies have been investigated the effect of test length (Fidalgo et al., 2000; Rogers & Swaminathan, 1993; Uttaro & Millsap, 1994). Thus, two levels of test length (i.e., 30 and 60 items) were considered in the study.

An item is flagged as DIF when people with the same latent ability but from different groups have an unequal probability of answering an item correctly. In CDM, DIF happens when item parameters (i.e., s_j and g_j) of the reference and focal groups are different. Previous studies have investigated the DIF amount and DIF type (i.e., uniform DIF and nonuniform DIF) (Hou et al., 2014; Liu et al., 2019; Sünbül, 2019; Zhang, 2006). Thus, seven levels of DIF type were manipulated in this study, including no DIF, uniform DIF, and nonuniform DIF. When the difference of item parameters between two groups was zero (i.e., $s_{Fj}-s_{Rj}=0$ and $g_{Fj}-g_{Rj}=0$), an item j has no DIF. When the difference of item parameters between two groups was not zero, an item has DIF. Similar to Hou et al. (2014) and Liu et al. (2019), the difference of item parameters between two groups was set to be 0.1. Uniform DIF occurs in the item j when ($s_{Fj}-s_{Rj}>0$ and $g_{Fj}-g_{Rj}<0$) or ($s_{Fj}-s_{Rj}<0$ and $g_{Fj}-g_{Rj}>0$), in which the item j favors the reference or the focal groups, respectively. Two uniform DIF types were ($s_{Fj}-s_{Rj}=+0.1$ and $g_{Fj}-g_{Rj}=-0.1$) or ($s_{Fj}-s_{Rj}=-0.1$ and $g_{Fj}-g_{Rj}=+0.1$). Nonuniform DIF occurs in the item j when the two item parameters favor different groups. For the first nonuniform DIF type ($s_{Fj}-s_{Rj}=+0.1$ and $g_{Fj}-g_{Rj}=0$), the item j favors the

reference group in slip parameter but doesn't favor any group in guessing parameter. For the second nonuniform DIF type ($s_{Fj} - s_{Rj} = 0$ and $g_{Fj} - g_{Rj} = -0.1$), the item j favors the reference group in guessing parameter but doesn't favor any group in slip parameter. For the third nonuniform DIF type ($s_{Fj} - s_{Rj} = +0.1$ and $g_{Fj} - g_{Rj} = +0.1$), the item j favors the reference group in slip parameter but favors the focal group in guessing parameter. For the fourth nonuniform DIF type ($s_{Fj} - s_{Rj} = -0.1$ and $g_{Fj} - g_{Rj} = -0.1$), the item j favors the focal group in slip parameter but favors the reference group in guessing parameter.

The DIF percentage has been investigated in previous studies (Fidalgo et al., 2000; Jin et al., 2018; Sünbül, 2019). To understand the effect of DIF percentage on DIF detection, 0%, 10%, 20%, 30%, and 40% DIF items were manipulated in this study. Many DIF detection methods have been used in CDMs (Hou et al., 2014; Sünbül, 2019; Zhang, 2006). As previous discussion, the MH and OR were considered in this study. The DIF method had five levels, including the MH, MH-P, MH-A, OR, and OR-P methods. The MH method was the MH with total scores as the matching variables. The MH-P method was the previous MH method with purification procedure. The MH-A method was the MH method with attribute patterns as the matching variables. The OR and OR-P method were the OR (Jin et al., 2018) method and OR with purification procedure, respectively.

To sum up, seven variables were manipulated: model (two levels: DINA and DINO), sample size for reference and focal groups (six levels: 500/500, 1000/500, 1000/1000, 2000/500, 2000/1000, and 2000/2000), item parameters (two levels: .25 and .75), test length (two levels: 30 and 60 items), DIF type (seven levels: no DIF, two uniform DIF types, and four nonuniform DIF types), DIF percentage (five levels: 0%, 10%, 20%, 30%, and 40%), and DIF method (five levels: MH, MH-P, MH-A, OR, and OR-P). Each condition was replicated 100 times to reduce sampling error. To evaluate the performance of DIF detection method, evaluation criteria included averaged type I error and averaged statistical power over 100 replications.

3 Results

The type I error for the DINA model in the no DIF condition list in Table 1. The type I error should be approximately 5% to meet the model expectations under each condition. According to Wang and Su (2004)'s suggestion, if the type I error was less than 4% or more than 6%, the type I error in Table 1 appeared in bold. In the no DIF condition, the type I error for the DINA model approximately met the model expectation under different sample size, test length, item parameter, and DIF methods. The DINO model in the no DIF condition performed similarly to the DINA model.

Owing to spatial limitations, the results of the DINA model in one uniform DIF type were shown in Figs. 1 and 2. When the uniform DIF type was ($s_{Fj} - s_{Rj} = 0.1$ and $g_{Fj} - g_{Rj} = -0.1$), the type I error and power of the five DIF methods for different DIF percentage and sample size in the DINA model were shown in Figs. 1

Table 1 Type I error for the DINA model in the no DIF condition

Sample size	Test length	Item parameter	MH	MH-P	MH-A	OR	OR-P
500/500	30	0.25	0.0433	0.0423	0.0413	0.0463	0.0463
		0.75	0.0403	0.0403	0.0413	0.0440	0.0467
	60	0.25	0.0407	0.0415	**0.0387**	0.0462	0.0468
		0.75	0.0410	**0.0392**	0.0402	0.0433	0.0437
1000/500	30	0.25	**0.0387**	**0.0367**	**0.0397**	0.0403	0.0413
		0.75	0.0430	**0.0400**	**0.0377**	0.0423	0.0437
	60	0.25	0.0423	0.0440	0.0428	0.0437	0.0445
		0.75	0.0410	0.0413	0.0422	0.0428	0.0438
1000/1000	30	0.25	0.0450	0.0453	0.0450	0.0470	0.0497
		0.75	0.0447	0.0440	0.0413	0.0450	0.0473
	60	0.25	0.0402	0.0410	0.0430	0.0418	0.0437
		0.75	0.0497	0.0498	0.0460	0.0487	0.0508
2000/500	30	0.25	0.0413	0.0413	**0.0397**	0.0453	0.0463
		0.75	0.0443	0.0453	0.0433	0.0450	0.0467
	60	0.25	0.0402	0.0410	0.0430	0.0418	0.0437
		0.75	0.0497	0.0498	0.0460	0.0487	0.0508
2000/1000	30	0.25	0.0413	0.0413	**0.0373**	**0.0383**	0.0423
		0.75	0.0440	0.0433	0.0450	0.0443	0.0457
	60	0.25	0.0422	0.0427	0.0417	0.0420	0.0437
		0.75	0.0428	0.0430	0.0405	0.0433	0.0437
2000/2000	30	0.25	0.0470	0.0473	0.0423	0.0467	0.0523
		0.75	0.0463	0.0457	0.0457	0.0480	0.0500
	60	0.25	0.0462	0.0468	0.0455	0.0430	0.0448
		0.75	0.0432	0.0450	0.0427	0.0427	0.0465

and 2. The DIF method performs well if its type I error meet the model expectations (Wang & Su, 2004), and the corresponding power is high. The type I error of the MH-A method in the DINA model approximately met the model expectation under different sample size and DIF percentage, and the MH-A method in the DINA model had the highest power among all other methods. As the sample size or DIF percentage increased, the type I error of the other methods increased and the power of the other methods decreased. The MH method performed the worst in terms of the inflated type I error. The results of the DINA model in the other uniform DIF condition (i.e., $s_{Fj} - s_{Rj} = -0.1$ and $g_{Fj} - g_{Rj} = +0.1$) were similar to those in

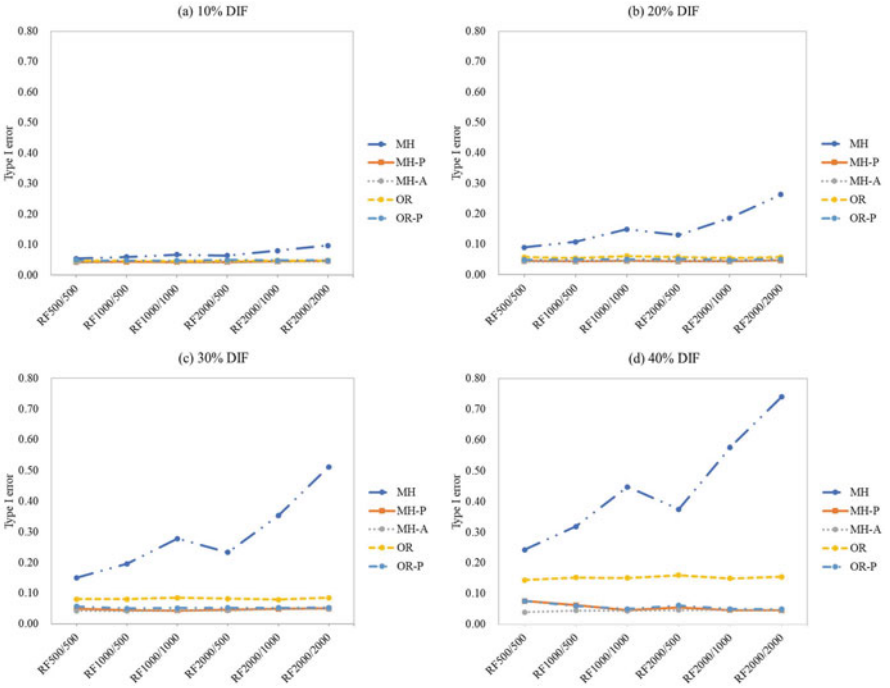


Fig. 1 Type I error of the DIF methods for different DIF percentage in DINA model when the uniform DIF type was ($s_{Fj} - s_{Rj} = 0.1$ and $g_{Fj} - g_{Rj} = -0.1$)

Figs. 1 and 2. Besides, the results of the DINO model were similar to those of the DINA model in Figs. 1 and 2.

When the nonuniform DIF type was ($s_{Fj} - s_{Rj} = 0.1$ and $g_{Fj} - g_{Rj} = 0$), the DINA model had lower type I error and power than the DINO model. When the nonuniform DIF type was ($s_{Fj} - s_{Rj} = 0$ and $g_{Fj} - g_{Rj} = -0.1$), the DINA model had higher type I error and power than the DINO model. When the nonuniform DIF type was ($s_{Fj} - s_{Rj} = 0.1$ and $g_{Fj} - g_{Rj} = 0.1$) or ($s_{Fj} - s_{Rj} = -0.1$ and $g_{Fj} - g_{Rj} = -0.1$), the DINA model had similar type I error and power compared to the DINO model.

4 Conclusions and Discussion

Several conclusions were drawn from this study. First, for the uniform DIF type, the DINA model performed similarly to the DINO model. Both the DINA and DINO models had higher power for the uniform DIF type than that for the nonuniform DIF types. For four nonuniform DIF types, the DINA model performed differently from the DINO model. Those findings were similar to the previous studies (Hou et

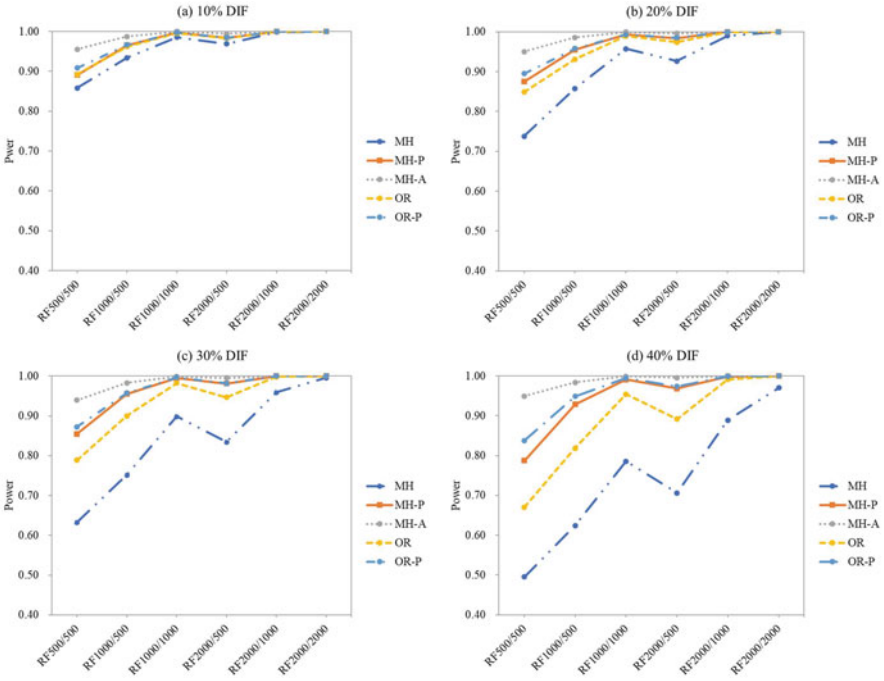


Fig. 2 Power of the DIF methods for different DIF percentage in DINA model when the uniform DIF type was ($s_{Fj} - s_{Rj} = 0.1$ and $g_{Fj} - g_{Rj} = -0.1$)

al., 2014; Liu et al., 2019). Second, generally speaking, the power was lower when the item parameters were set to be .75, which item bank was low quality. This was similar to Hou et al. (2014). However, it had higher power when the item parameters were set to be .75 (i.e., low-quality item pool) in few DIF conditions. This would need further investigation. Third, test length had less effect on DIF detection than expectation. Fourth, the type I error of the MH, MH-A, and MH-P methods was increased with the sample size increased. The type I error of the OR and OR-P methods was less affected by the sample size. The power of all the DIF detection methods was increased with the simple size increased. Fifth, the type I error was inflated when the DIF percentage was increased. Those findings were similar to the previous studies (Fidalgo et al., 2000; Jin et al., 2018; Sünbül, 2019). Sixth, the MH-A method had the best performance in terms of the lowest type I error and highest power. This is because using the attribute patterns is the better than the total scores as the matching variables in CDM. The OR-P method performed well in large sample size, and the power was less affected when DIF percentage was high.

Some limitations were found in this study. First, this study set the fixed values for the item parameters, which is similar to Hou et al. (2014); however, some other studies used the uniform distribution to generated item parameters (Zhang, 2006). In practice, the uniform distribution can be used to generated the item parameters.

Second, this study was conducted in DINA and DINO models. Previous studies have used numerous CDMs. Investigating the performance of the five DIF methods in different CDMs is of great interest. Third, in this study, the examinees' attribute patterns were generated according to Zhang (2006)'s method. Different methods would be used to generate attribute patterns, and then might have different impact on the results. Fourth, this study was conducted when attributes have a nonhierarchical relationship. In practice, attributes might have a hierarchical relationship, meaning some are a prerequisite for the presence of others. It is interesting to investigate the performance of DIF detection methods when attributes have a hierarchical relationship.

References

Chiu, C.-Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification, 30*(2), 225–250. https://doi.org/10.1007/s00357-013-9132-9

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199. https://doi.org/10.1007/s11336-011-9207-7

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353. https://doi.org/10.1007/BF02295640

Fidalgo, A. M., Mellenbergh, G. J., & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel–Haenszel procedures. *Methods of Psychological Research, 5*(3), 43–53.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333–352.

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

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the Mantel-Haenszel procedure. *ETS Research Report Series, 1986*(2), i–24. https://doi.org/10.1002/j.2330-8516.1986.tb00186.x

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Lawrence Erlbaum Associates, Inc.

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement, 51*(1), 98–125. https://doi.org/10.1111/jedm.12036

Jin, K. Y., Chen, H. F., & Wang, W. C. (2018). Using odds ratios to detect differential item functioning. *Applied Psychological Measurement, 42*(8), 613–629. https://doi.org/10.1177/0146621618762738

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.

Li, F. (2008). *A modified higher-order DINA model for detecting differential item functioning and differential attribute functioning*. Unpublished doctoral dissertation, University of Georgia.

Li, X., & Wang, W.-C. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement, 52*(1), 28–54. https://doi.org/10.1111/jedm.12061

Liu, Y., Yin, H., Xin, T., Shao, L., & Yuan, L. (2019). A comparison of differential item functioning detection methods in cognitive diagnostic models. *Frontiers in Psychology, 10*, 1137. https://doi.org/10.3389/fpsyg.2019.01137

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *JNCI: Journal of the National Cancer Institute, 22*(4), 719–748. https://doi.org/10.1093/jnci/22.4.719

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 87–212.

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

Morrison, D. F. (1967). *Multivariate statistical methods*. McGraw-Hill.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2), 105–116. https://doi.org/10.1177/014662169301700201

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007). The fusion model skills diagnosis system. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 275–318). Cambridge University Press.

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

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159–194. https://doi.org/10.1007/BF02294572

Sünbül, S. (2019). Investigating differential item functioning in DINA model. *International Journal of Progressive Education, 15*, 174–186. https://doi.org/10.29329/ijpe.2019.203.13

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

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

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305. https://doi.org/10.1037/1082-989x.11.3.287

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement, 18*(1), 15–25. https://doi.org/10.1177/014662169401800102

Wang, W.-C., & Su, Y.-H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education, 17*(2), 113–144. https://doi.org/10.1207/s15324818ame1702_2

Zhang, W. (2006). *Detecting differential item functioning using the DINA model*. Unpublished doctoral dissertation, University of North Carolina.

Effect of Within-Group Dependency on Fit Statistics in Mokken Scale Analysis in the Presence of Two-Level Test Data

Letty Koopman

Abstract Investigating model fit is essential for determining measurement properties of tests and questionnaires. Mokken scale analysis (MSA) consists of a selection of methods to investigate the fit of nonparametric item response theory models. Existing MSA methods assume a simple random sample, which is violated in two-level test data (i.e., test data of clustered respondents). This chapter discusses the methods manifest monotonicity, conditional association, and manifest invariant item ordering, and investigates the effect of within-group dependency on the point estimate and variability of their statistics. Results showed that fit statistics may be safely used in the presence of within-group dependency, giving appropriate results for sets of items that either did or did not violate assumptions. Implications for practice are discussed.

Keywords Conditional association · Manifest invariant item ordering · Manifest monotonicity · Model fit · Mokken scale analysis

1 Introduction

Mokken scale analysis (MSA) consists of a selection of methods to investigate the fit of nonparametric item response theory models (IRT; see, e.g., Mokken, 1971; Sijtsma and Molenaar, 2002; Sijtsma and Van der Ark, 2017). Let θ denote a latent variable. Let X_i denote a binary latent variable with realization x_i that takes on value 1 if an item is endorsed or correctly answered, and 0 otherwise. Let $P(X_i = 1|\theta)$ denote the item-response function of item i, which is the probability of correctly

L. Koopman (✉)
Centre for Educational Measurement, University of Oslo , Oslo, Norway

Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, The Netherlands
e-mail: V.E.C.Koopman@cemo.uio.no

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_20

scoring item i as a function of the latent variable. Four main assumptions of IRT models are

1. Unidimensionality: Latent variable θ is unidimensional
2. Local independence: The item scores are independent given the respondent's value on θ
3. Monotonicity: Item-response function $P(X_i = 1|\theta)$ is monotonically nondecreasing.
4. Invariant item ordering: Item-response functions $P(X_i = 1|\theta)$ and $P(X_j = 1|\theta)$ do not intersect for all $i \neq j$ and for all θ.

Item response models that limit themselves to only the first three assumptions (monotone homogeneity model) or first four assumptions (double monotonicity model) can be classified as nonparametric IRT models, because they pose no distributional assumptions on the latent variable and the item-response function (e.g., Mokken, 1971; Sijtsma & Molenaar, 2002).

Different combinations of these assumptions imply various observable data characteristics that can be used to investigate fit of nonparametric IRT models, including manifest monotonicity (Junker, 1993; Sijtsma and Hemker, 2000), conditional association (Holland & Rosenbaum, 1986; Straat et al., 2016), and manifest invariant item ordering (Ligtvoet et al., 2010, 2011). The suggested practice to identify scales using MSA is to first investigate scalability using scalability coefficients, after which local independence, monotonicity, and invariant item order can be investigated using the properties manifest monotonicity, conditional association, and manifest monotonicity, respectively (Sijtsma & Van der Ark, 2017).

The estimation method for these procedures assume the data are obtained by means of a simple random sampling design. However, this assumption is often violated, for example in two-stage sampling design in which first schools are sampled, after which students within the sampled school are sampled. Two-stage sampling designs are standard in large-scale international assessments such as TIMSS, PIRLS, and PISA (e.g., Joncas & Foy, 2011; Karjalainen & Laaksonen, 2008). Such sampling designs often lead to within-group dependency in the data, where the item scores of respondents within the same group are more related than item scores of respondents across different groups.

Snijders (2001) proposed nonparametric IRT models for two-level test data, and Koopman (2022, Chapter 8) showed that, in the population, these models imply the same observable properties as their one-level counterparts. However, within-group dependency can substantially affect statistical analyses in data samples, requiring methods that acknowledge and take into account the nested structure of the data. For example, Koopman et al. (2020, 2022) showed the quality of items and the total scale may be overestimated when the nested structure is ignored, and provided alternative methods. Currently, it is unknown to which degree the fit statistics are affected by within-group dependency.

This chapter first provides an outline on the properties manifest monotonicity, conditional association, and manifest invariant item ordering, and discusses the currently implemented methods for evaluating them in data. Next, a simulation

study is presented that evaluated the effect of a nested data structure on the fit statistics provided by these methods for different degrees of within-group dependency (from no dependency to extreme dependency) and for data simulated with and without items that violated assumptions.

2 Observable Properties

Various observable properties are implied by the four main assumptions of unidimensionality, local independence, monotonicity, and invariant item ordering. Here I discuss three of them, manifest monotonicity, conditional association, and manifest invariant item ordering.

2.1 Manifest Monotonicity

Let $X_{(i)} = \sum_{j \neq i}^{I} X_j$ denote the rest score of item i, which is the sum across all items except item i. Manifest monotonicity means that $P(X_i = 1|X_{(i)})$ is nondecreasing in $X_{(i)}$. For dichotomous items, unidimensionality, local independence, and monotonicity imply manifest monotonicity (Junker & Sijtsma, 2000). Manifest monotonicity is used to investigate the assumption of (latent) monotonicity. For each item the rest score $X_{(i)}$ is computed, after which the proportion of respondents scoring 1 on item X_i is calculated, estimating $P(X_i = 1|X_{(i)})$. Manifest monotonicity is violated if $P(X_i = 1|X_{(i)} = l) > P(X_i = 1|X_{(i)} = l + 1)$. If rest score groups are too small according to a set criterion (*minsize*), consecutive rest score groups are joined. The number of rest score groups that are evaluated determines the number of active comparisons, and these comparisons check for violations of manifest monotonicity. To avoid evaluating violations that are too small for practical meaning, a set criterion (*minvi*, default = 0.03) determines when violations are large enough to be counted and tested for significance. The violation is tested for significance using a normal approximation to the hypergeometric distribution. Diagnostic value *crit* combines evidence of various statistics to give an indication of whether an item violates the assumptions. Table 1 provides an overview of the summary statistics that are provided when investigating manifest monotonicity in R (Van der Ark, 2007) or in MSP5 (Molenaar & Sijtsma, 2000).

2.2 Conditional Association

Let vector $\mathbf{X} = (X_1, X_2, \ldots, X_I)$ contain the I item scores X_i, and divide $\mathbf{X}$ into two mutually exclusive vectors $\mathbf{Y}$ and $\mathbf{Z}$. Let g_1 and g_2 be nondecreasing functions, h any function, and σ the population covariance. Conditional association is defined as

Table 1 Statistics for manifest monotonicity

Statistic	Description
#ac	The number of active comparisons
#vi	The number of violations of manifest monotonicity
#vi/#ac	The average number of violations per comparison
maxvi	The largest violation of manifest monotonicity
sum	The sum of violations of manifest monotonicity
sum/#ac	The average violation per active comparison
zmax	The maximum test statistic
#zsig	The number of violations that are significantly greater than zero
crit	The crit value

Note. Only violations are counted that exceed $minvi$

Table 2 Statistics for conditional association

Statistic	Description
$W_{ij}^{(1)}$	Identifying positive local dependence between items i and j
$W_{i}^{(2)}$	Identifying whether item i to likely be in a positively dependent item pair
$W_{ij}^{(3)}$	Identifying negative local dependence between items i and j

$$\sigma[g_1(\mathbf{Y}), g_2(\mathbf{Z})|h(\mathbf{Z}) = \mathbf{z}] \geq 0 \tag{1}$$

(Holland & Rosenbaum, 1986, Definition 3.4). Unidimensionality, local independence, and monotonicity imply conditional association (Holland & Rosenbaum, 1986, Theorem 6). Let $X_{(ij)}$ denote the rest score on all items except item i and j. Let $\mathbf{Y} = (X_i, X_j)$ and $\mathbf{Z}$ the scores on the remaining items. In that case, conditional association implies $\sigma(X_i, X_j) \geq 0$ (i.e., $\mathbf{Z}$ is ignored), $\sigma(X_i, X_j|X_k = x) \geq 0$ (i.e., conditioning on item X_k), and $\sigma(X_i, X_j|X_{(ij)} = l) \geq 0$ (i.e., conditioning on the rest score). Straat et al. (2016) proposed three statistics that use conditional association to evaluate the local independence assumption, presented in Table 2.

2.3 Manifest Invariant Item Ordering

Manifest invariant item ordering means that if for $i < j$ it holds that $P(X_i = 1) \geq P(X_j = 1)$ (i.e., item i is in general easier or more popular than item j), then $P(X_i = 1|R_{(ij)} = l) \geq P(X_j = 1|R_{(ij)} = l)$ for all l and all $i < j$. Unidimensionality, local independence, monotonicity, and invariant item ordering (i.e., the four main assumptions) imply manifest invariant item ordering (Ligtvoet et al., 2011). Table 3 provides an overview of the summary statistics that are provided when investigating manifest invariant item ordering, which is similar to method manifest monotonicity.

Table 3 Statistics for manifest invariant item ordering

Statistic	Description
#ac	The number of active comparisons
#vi	The number of violations of manifest monotonicity
#vi/#ac	The average number of violations per active comparison
maxvi	The largest violation of manifest monotonicity
sum	The sum of violations of manifest monotonicity
sum/#ac	The average violation per active comparison
zmax	The maximum test statistic
#zsig	The number of violations that are significantly greater than zero
crit	The crit value

Note. Only violations are counted that exceed $minvi$

3 Simulation Study

The effect of within-group dependency on the fit statistics were evaluated in a Monte Carlo simulation study[1] (see, e.g., Morris et al., 2019). The goal was to investigate the effect of within-group dependency on the estimated fit statistics, for test data that complied with the assumptions or test data that violated the assumptions.

3.1 Method

Data were generated for 50 groups consisting of 50 respondents, using a two-dimensional two-parameter logistic model to allow for violating assumptions. Note that without violated assumptions, this model is a parametric special case of the monotone homogeneity model (Van der Ark, 2001) For each respondent scores were sampled to 10 dichotomous items, indexed i ($i = 1, 2, \ldots, 10$), for 50 groups, indexed s ($s = 1, 2, \ldots, 50$), each consisting of a unique set of 50 respondents, indexed r ($r = 1, 2, \ldots, 50$). Let α_i and β_i denote the discrimination and difficulty of item i, respectively. Let α_i^* denote the second discrimination parameter in the case of local dependency. Let θ_{sr} denote the latent variable of interest for respondent r within group s, and θ_{sr}^* the latent nuisance variable for that respondent in the case of local dependency. Then, the probability that respondent r in group s scores 1 on item i is defined as

$$P(X_i = 1|\theta_{sr}) = \frac{exp[\alpha_i(\theta_{sr} - \beta_i) + \alpha_i^*(\theta_{sr}^* - \beta_i)]}{1 + exp[\alpha_i(\theta_{sr} - \beta_i) + \alpha_i^*(\theta_{sr}^* - \beta_i)]}. \quad (2)$$

Data were simulated across $Q = 1000$ replications

[1] Syntax files are available to download from the Open Science Framework: https://osf.io/hfn6j/.

3.1.1 Independent Variables

Two independent variables were included.

Within-Group Dependency Within-group dependency had three degrees: independent, medium, and extreme. To manipulate within-group dependency value θ_{sr} was split in $\theta_{sr} = \gamma_s + \delta_{sr}$, in which γ_s is a group-specific value and δ_{sr} a respondent-specific value. Values γ_s and δ_{sr} were sampled separately and independently.

For the independent condition, γ_s was set to zero and $\theta_{sr} = \delta_{sr} \sim N(0, 1)$. Hence, there was no within-group dependency, meaning that there is no group effect. Essentially, this type of test data is equal to test data obtained by a simple random sample.

For the medium condition, both γ_s and $\delta_{sr} \sim N(0, \sigma^2 = 0.5)$, meaning that the group- and respondent-effect is equal, usually leading to a substantial amount of within-group dependency.

For the extreme condition, δ_{sr} was set to zero and $\theta_{sr} = \gamma_s \sim N(0, 1)$. Hence, there was a maximum amount of within-group dependency, meaning that there is no individual respondent effect. This is the most extreme type of within-group dependency given the item parameters and item response theory model. Note that for all conditions, the variance of $\theta_{sr} = 1$. The latent nuisance variable $\theta^*_{sr} \sim N(0, 1)$ and was sampled similarly as θ_{sr} in the different conditions (i.e., by sampling only γ^*_s, only δ^*_{sr}, or both), although the latent variables were unrelated.

Violation of Assumptions This variable had two conditions: Absence and presence of violated assumptions. This variable was manipulated by changing the item parameters. Data were simulated in the absence of violated assumptions using the following item parameters. Item discrimination $\alpha_i = 1$ was equal for all i. Item difficulty β_i had equidistant values between -1 and 1. $\alpha^*_i = 0$, hence there was no local dependency. Figure 1 shows the item-response functions for the conditon without violated assumptions.

Data were simulated in the presence of violated assumptions using the following item parameters. Item discrimination took on values $(-0.5, 0.5, 1, 1, 1, 1, 1, 1, 1, 2)$ for item $i = 1, \ldots, 10$, respectively. Item 1 violates the monotonicity assumption and item 1, 2, and 10 violate the assumption of invariant item ordering. Item difficulty value $\beta_1 = -1$, and the remaining items had equidistant values between -1 and 1, hence $\beta_1 = \beta_2$. Items 1 and 2 violate the assumption of invariant item ordering. $\alpha_i = 0$ for item $i = 1$ to 8 and $\alpha_9 = \alpha_{10} = 0.5$, causing local dependency between item 9 and 10, violating the local independence assumption. Hence, item 1, 2, 9, and 10 were the items violating assumptions. (i.e., removing these items gives an item set that complies with the assumptions). Figure 2 shows the item-response functions for the conditon with violated assumptions.

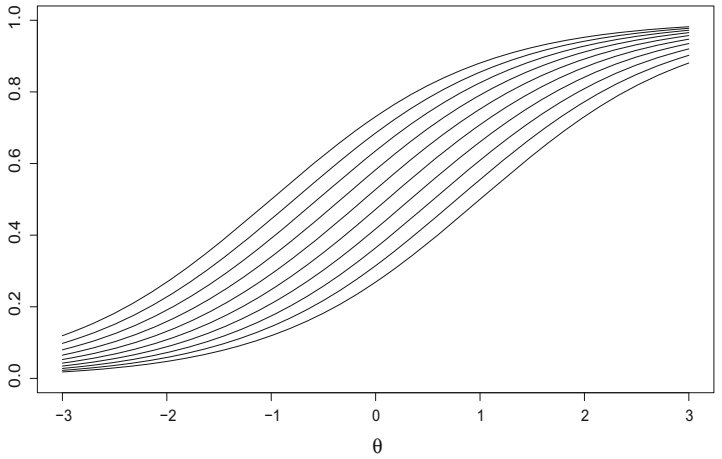


Fig. 1 Item response functions ($P(X_i = 1|\theta)$) of the ten items that did not violate assumptions

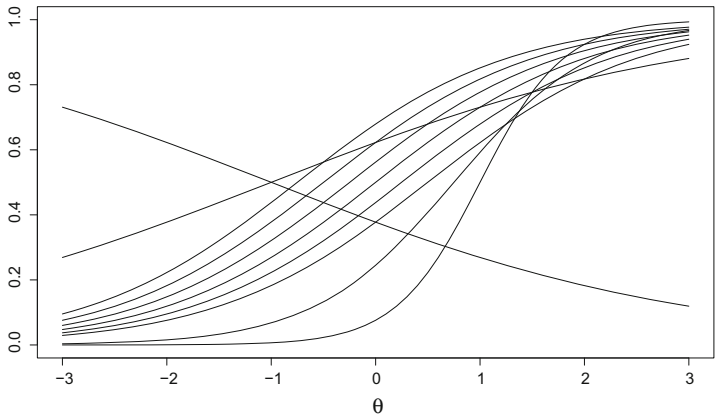


Fig. 2 Item response functions ($P(X_i = 1|\theta)$) of the ten items that violated assumptions

3.1.2 Dependent Variables

Accuracy and efficiency were computed for all statistics presented in Tables 1, 2, and 3.

Accuracy In general, all computed statistics are assumed accurate and optimal in the independent condition (i.e., no within-group dependency), as these test data have the same dependency structure as test data obtained using a simple random sampling design. Hence, the point estimates were compared to this condition

Efficiency The sampling fluctuation of a fit statistic is the standard deviation of the statistic across all replications. Let T denote a fit statistic. Then, the sampling fluctuation of T, denoted $S(T)$, is computed as

$$S(T) = \sqrt{\frac{1}{Q}\sum_{q=1}^{Q}(T_q - \overline{T})^2}. \tag{3}$$

The sampling fluctuation is a measure of efficiency.

3.1.3 Statistical Analyses

Fit statistics were computed using the default settings of the following functions from the R-package `mokken`. Let `R>` denote the R prompt and let `data` denote the particular dataset at hand.

```
R> # Load R-package mokken:
R> library(mokken)
R> # Manifest monotonicity:
R> summary(check.monotonicity(data))
R> # Conditional association:
R> check.ca(data, TRUE)$Index
R> # Manifest invariant item ordering:
R> summary(check.iio(data, item.selection = FALSE))$
R+         item.summary
```

Differences in point estimates and variability greater than 0.05 between the independent and other conditions are discussed. For all conditions, output will be averaged across items. Intraclass correlations (ICCs) will be computed for the each condition to reflect the degree of within-group dependency.

3.1.4 Hypotheses

In general, within-group dependency was expected to have no effect on the accuracy of the computed statistics, but the sampling fluctuation was expected to increase for higher levels of dependency, possibly leading to a higher number of significant violations. In general, statistics in all methods are higher when items violate assumptions, hence, their values were expected to be higher for conditions with violated assumptions compared to conditions without violated assumptions. No interaction effect between within-group dependency and assumption violations were expected.

4 Results

The ICCs reflected the intended substantial increase of the within-group dependency for the different conditions. In both independent conditions $\overline{ICC} = 0.000$ ($S = 0.004$). In the conditions without violated assumptions, $\overline{ICC} = 0.327$ ($S = 0.046$) and $\overline{ICC} = 0.660$ ($S = 0.045$) for the medium and extreme within-group dependency conditions, respectively, versus $\overline{ICC} = 0.278$ ($S = 0.043$) and $\overline{ICC} = 0.562$ ($S = 0.049$) in the conditions with violated assumptions, showing that the presence of violated assumptions decreased the within-group dependency slightly.

4.1 Manifest Monotonicity

For conditions without violated assumptions, the average #ca and crit were higher for higher within-group dependency conditions (Table 4, columns one to three). In addition, the sampling fluctuation of #ca, zmax, and crit was higher for higher levels of within group dependency.

For conditions with violated assumptions, the average #ac, #vi, #zsig, and crit were slightly lower for higher within-group dependency conditions. In addition, the sampling fluctuation of #ac, #vi, zmax, #zsig, and crit was higher for higher levels of within-group dependency (Table 4, columns four to six).

4.2 Conditional Association

For conditions with and without violated assumptions, the estimated W_i^2 and its standard error were higher for higher within-group dependency conditions. The

Table 4 Magnitude and sampling fluctuation (in parentheses) of manifest monotonicity statistics

	Without violations			With violations		
	Independent	Medium	Extreme	Independent	Medium	Extreme
#ac	19.21 (2.20)	19.76 (3.20)	19.64 (3.63)	16.11 (1.16)	15.57 (1.98)	15.38 (2.58)
#vi	0.00 (0.07)	0.01 (0.09)	0.01 (0.10)	1.78 (0.28)	1.70 (0.34)	1.65 (0.39)
#vi/#ac	0.00 (0.00)	0.00 (0.00)	0.00 (0.01)	0.09 (0.01)	0.09 (0.01)	0.09 (0.01)
maxvi	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.03 (0.01)	0.03 (0.01)	0.03 (0.01)
sum	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.23 (0.05)	0.22 (0.06)	0.21 (0.07)
sum/#ac	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)
zmax	0.00 (0.07)	0.01 (0.10)	0.01 (0.12)	0.76 (0.15)	0.77 (0.16)	0.76 (0.21)
#zsig	0.00 (0.00)	0.00 (0.01)	0.00 (0.03)	1.50 (0.23)	1.42 (0.27)	1.37 (0.31)
crit	0.10 (1.49)	0.21 (2.23)	0.28 (2.74)	42.85 (4.61)	42.61 (5.19)	41.96 (6.57)

Table 5 Magnitude and sampling fluctuation (in parentheses) of conditional association statistics

	Without violations			With violations		
	Independent	Medium	Extreme	Independent	Medium	Extreme
W_{ij}^1	0.02 (0.04)	0.02 (0.04)	0.04 (0.09)	3.19 (0.21)	3.20 (0.23)	3.21 (0.28)
W_i^2	18.52 (2.01)	18.71 (2.14)	19.27 (2.57)	23.86 (2.46)	24.00 (2.54)	24.34 (2.81)
W_{ij}^3	2.06 (0.69)	2.08 (0.69)	2.14 (0.71)	2.65 (0.68)	2.67 (0.69)	2.70 (0.71)

Table 6 Magnitude and sampling fluctuation (in parentheses) of manifest invariant item ordering statistics

	Without violations			With violations		
	Independent	Medium	Extreme	Independent	Medium	Extreme
#ac	51.72 (1.62)	50.31 (2.81)	48.89 (3.60)	42.87 (1.16)	42.74 (1.96)	42.34 (2.77)
#vi	0.16 (0.39)	0.14 (0.38)	0.14 (0.37)	3.92 (1.54)	3.80 (2.01)	3.62 (2.30)
#vi/#ac	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	0.09 (0.03)	0.09 (0.04)	0.08 (0.05)
maxvi	0.01 (0.01)	0.00 (0.01)	0.00 (0.01)	0.25 (0.04)	0.24 (0.05)	0.23 (0.08)
sum	0.01 (0.02)	0.01 (0.02)	0.00 (0.01)	0.62 (0.22)	0.59 (0.32)	0.56 (0.39)
sum/#ac	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.00)	0.01 (0.01)	0.01 (0.01)
zmax	0.18 (0.44)	0.16 (0.44)	0.16 (0.42)	6.77 (1.12)	6.76 (1.58)	6.70 (2.34)
#zsig	0.02 (0.13)	0.02 (0.13)	0.01 (0.11)	3.21 (1.32)	3.09 (1.75)	2.94 (2.03)
crit	3.04 (7.71)	2.80 (7.61)	2.64 (7.20)	120.37 (21.33)	117.71 (30.89)	114.21 (40.56)

W_{ij}^3 was on average slightly higher for higher within-group dependency conditions, whereas for W_{ij}^1 only the sampling fluctuation was higher in higher within-group dependency conditions (Table 5).

4.3 Manifest Invariant Item Ordering

For the conditions without violated assumptions, the average #ac and crit value were lower for higher levels of within-group dependency (Table 6, columns one to three). The sampling fluctuation was higher for #ac, but lower for crit. For conditions with violated assumptions, the average #ac, #vi, sum, zmax, #zsig, and crit values were lower for higher within-group dependency conditions. In addition, the sampling fluctuation of #ac, #vi, sum, zmax, #zsig, and crit values were higher for higher levels of within-group dependency (Table 6, columns four to six).

5 Discussion

This chapter investigated the effect of within-group dependency on the accuracy and efficiency of various fit statistics provided by methods manifest monotonicity,

conditional association, and manifest invariant item ordering. Results showed that, in general, the magnitude of all fit statistics was very similar across the different within-group dependency conditions. In addition, the statistics that should identify violations of the assumptions were substantially higher in the conditions with violated assumptions.

For manifest monotonicity and manifest invariant item ordering, the number of active comparisons and number of violations fluctuated slightly more for higher within-group dependency conditions. However, when there are more comparisons, there are more possibilities for violations to occur. Therefore, the average number of violations per comparison is arguably more interesting, and that was unaffected by the level of within-group dependency, as was the average violation per comparison. Hence, the average number of violations as well as the average violation per comparison were accurately and efficiently estimated regardless of the degree of within-group dependency. The maximum z value, number of significant violations, and crit value fluctuated slightly more across samples for higher within-group dependency conditions. Hence, these values are less efficiently estimated in samples with larger within-group dependency, although the effect was negligible.

For the three conditional association coefficients the magnitude and sampling fluctuation was similar across within-group dependency conditions. Of these fit statistics, statistic W_i^2 was to the highest degree affected, but this may be due to the fact that its general magnitude and sampling fluctuation was substantially larger compared to the other two statistics. To conclude, the results suggest that fit statistics may be used safely in the presence of within-group dependency, giving appropriate results for items that did or did not violate assumptions.

References

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics, 14*(4), 1523–1543.

Joncas, M., & Foy, P. (2011). Sample design in TIMSS and PIRLS. *Methods and Procedures in TIMSS and PIRLS*, 1–21.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics, 21*(3), 1359–1378.

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24*(1), 65–81.

Karjalainen, T., & Laaksonen, S. (2008). PISA 2006 sampling and estimation. *Ministry of Education Publications 2008, 44*, 231–236.

Koopman, L. (2022). *Nonparametric Item Response Theory for Multilevel Test Data*. [Doctoral thesis, University of Amsterdam].

Koopman, L., Zijlstra, B. J. H., & Van der Ark, L. A. (2020). Standard errors of two-level scalability coefficients. *British Journal of Mathematical and Statistical Psychology, 73*(2), 213–236.

Koopman, L., Zijlstra, B. J. H., & Van der Ark, L. A. (2022). A two-step, test-guided Mokken scale analysis, for nonclustered and clustered data. *Quality of Life Research, 31*(1), 25–36.

Ligtvoet, R., Van der Ark, L. A., Bergsma, W. P., & Sijtsma, K. (2011). Polytomous latent scales for the investigation of the ordering of items. *Psychometrika, 76*(2), 200–216.

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*(4), 578–595.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Mouton.

Molenaar, I. W., & Sijtsma, K. (2000). *User's manual MSP5 for Windows*. IEC ProGAMMA.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine, 38*(11), 2074.

Sijtsma, K., & Hemker, B. T. (2000). A taxonomy of IRT models for ordering persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics, 25*(4), 391–415.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Sage.

Sijtsma, K., & Van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology, 70*(1), 137–158.

Snijders, T. A. B. (2001). Two-level non-parametric scaling for dichotomous data. In Boomsma, A., van Duijn, M. A. J., & Snijders, T. A. B. (Eds.), *Essays on item response theory* (pp. 319–338). Springer.

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2016). Using conditional association to identify locally independent item sets. *Methodology, 12*(4), 117–123.

Van der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement, 25*(3), 273–282.

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*(11), 1–19.

Regularized Robust Confidence Interval Estimation in Cognitive Diagnostic Models

Candice Pattisapu Fox and Richard M. Golden

Abstract Although prior work has investigated the effects of model misspecification on inference in Cognitive Diagnostic Models (CDMs) (e.g., Liu et al., Multivar Behav Res, 1–20, 2021), few studies have explored how regularization methods might impact issues of local identifiability and the quality of statistical inference in the presence of model misspecification. In the present study, Tatsuoka's (George et al., J Stat Software 74(2), 24, 2016; Tatsuoka, Analysis of errors in fraction addition and subtraction problems. Final report for NIE-G-81-0002, 1984) 15 question fraction-subtraction data set (n = 536) was fit to a five latent skill DINA CDM using a uniform attribute profile distribution. A Gaussian prior was introduced to regularize the DINA CDM using MAP estimation (Ma and Jiang, Appl Psychol Meas 45(2):95–111, 2021). Next, parametric bootstrap data sets were generated from the fitted model and fit to both the original model and a misspecified version of the original model. By including an informative Gaussian prior, confidence interval coverage estimation performance was shown to improve for the Robust covariance matrix estimator but not for the Hessian and OPG covariance matrix estimators.

Keywords Cognitive diagnostic models · Regularization · Misspecification · MAP estimation · Robust covariance matrix

1 Introduction

Cognitive Diagnostic Models (CDMs) are restricted, parameterized latent class models of assessments which predict student correct item response probability as a function of latent skill mastery for a particular exam item. Although CDMs have tremendous potential for improving the quality of formative assessments in classrooms, their full potential remains underutilized due to challenges associated

C. Pattisapu Fox (✉) · R. M. Golden
Cognitive Informatics and Statistics Lab, School of Behavioral and Brain Sciences (GR4.1), University of Texas at Dallas, Richardson, TX, USA
e-mail: candice.pattisapu@utdallas.edu; golden@utdallas.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_21

with smaller examinee populations relative to many high-stakes assessment efforts (Sessoms & Henson, 2018; Paulsen & Valdivia, 2022). In such cases, statistical regularities in the environment are not sufficient to support unique identification of parameter estimates and this process can be further influenced by the presence of model misspecification.

It is well-known that a Maximum Likelihood Estimate (MLE) has an asymptotic Gaussian distribution which, in turn, supports the construction of confidence intervals and hypothesis tests. In particular, if the probability model is correctly specified, the asymptotic Gaussian distribution of the MLE has a mean given by the true parameter value and a particular positive definite covariance matrix (e.g., Huber, 1967; White, 1982; Golden, 2020). In this special case of correct specification, the MLEs are asymptotically Gaussian with a covariance matrix which may be estimated using either the Hessian Covariance Matrix (derived from second derivatives of the likelihood function) or the Outer Product Gradient (OPG) Covariance Matrix (derived from first derivatives of the likelihood function and sometimes referred to as the Fisher Information Matrix or Cross-Product Matrix).

If the model is misspecified, then the maximum likelihood estimates can still be shown to have an asymptotic Gaussian distribution centered at the parameter values which minimize the cross-entropy of the fitted model relative to the distribution which generated the observed data. The asymptotic MLE covariance matrix in this case is called the Robust covariance matrix and is computed by combining the OPG and Hessian covariance matrix formulas (e.g., Huber, 1967; White, 1982; Golden, 2020) This approach also requires that both the OPG and Hessian covariance matrices are positive definite. This theoretical reason for using the Robust covariance matrix estimation methodology relative to the OPG and Hessian covariance matrix estimators when model misspecification is present has been also empirically supported in CDM simulation studies (e.g., Liu et al., 2021).

In a recent simulation study, Liu et al. (2021) have commented that the estimated accuracy of standard errors of parameter estimates appears to be affected in overspecified CDMs. Such situations may correspond to cases where the OPG or Hessian covariance matrix estimator is singular or near-singular. MAP (Maximum A Posteriori) estimation is an extension of ML estimation which provides opportunities for introducing prior knowledge to possibly improve identifiability and statistical inference quality (Mislevy, 1986; Maris, 1999; Golden, 2020; Ma and Jiang, 2021).

It can be shown (see discussion near Eq. 5) that a MAP estimation methodology with a Gaussian prior can be used to ensure at least that the Hessian covariance matrix estimator is positive definite for finite sample sizes. We refer to situations where the Gaussian prior keeps the magnitude of the Hessian covariance matrix estimator from becoming excessively large as “high regularization situations”. When model misspecification is present, however, this regularization strategy does not necessarily guarantee that the OPG covariance matrix estimator will be positive definite (Golden, 2022).

The objective of this study is to investigate the effects of model misspecification and regularization on CDM confidence interval coverage probabilities calculated

using the Hessian, Robust, and OPG covariance matrix estimation methods. Using a simulation study methodology, the parameters for correctly specified and misspecified models will be estimated using MAP estimation with a Gaussian prior. The magnitude of the covariance matrix of the Gaussian prior will also be manipulated for the purpose of contrasting low and high regularization situations.

2 Mathematical Theory

2.1 Cognitive Diagnostic Model Specification

2.1.1 Data Model

Assume the *question bank* consists of J questions. Let $\mathbf{x}_i$ be a J-dimensional binary vector whose jth element, x_{ij}, is equal to one if and only if examinee i correctly answers the jth question in the question bank. Let J-dimensional binary vector $\mathbf{x}_i$ denote the *response vector* for the ith examinee. Assume the *observed data* $\mathbf{x}_1, \ldots \mathbf{x}_n$ is a realization of a stochastic sequence of independent and identically distributed random vectors $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n$.

2.1.2 Probability Model of Data Generating Process

Let $\boldsymbol{\alpha}_i$ be a K-dimensional binary vector whose kth element is equal to one if and only if examinee i demonstrates mastery of latent skill k. Define the *latent skill profile* for the ith examinee as the K-dimensional binary vector $\boldsymbol{\alpha}_i$. Assume the responses for the ith examinee are denoted by $\mathbf{x}_i$ and the latent attribute skill profile for the ith examinee, $\boldsymbol{\alpha}_i$, are both observable. Let $\mathbf{q}_j = [q_{j,1}, \ldots, q_{j,K}]$ denote the jth row of the $\mathbf{Q}$ matrix whose element in row j and column k specifies that the kth skill is relevant for answering the jth question.

Let $\mathcal{S}(u)$ be defined such that $\mathcal{S}(u) = 1/(1 + \exp(-u))$. Let $\boldsymbol{\beta}_j = [\beta_{j,1}, \beta_{j,2}]^T$ be a two-dimensional column vector which consists of parameters specific to the jth question. The probability that the ith examinee correctly answers the jth question is given by the formula:

$$p\left(x_{ij} = 1 | \boldsymbol{\alpha}_i, \boldsymbol{\beta}_j, \mathbf{q}_j\right) = \mathcal{S}\left(\boldsymbol{\beta}_j \boldsymbol{\psi}(\boldsymbol{\alpha}_i, \mathbf{q}_j)\right).$$

A reparameterized DINA-type CDM (Henson et al., 2009) may be implemented by choosing $\boldsymbol{\psi}$ such that: $\boldsymbol{\psi}(\boldsymbol{\alpha}, \mathbf{q}_j) = \left[\left(\prod_{k=1}^{K} \alpha_k^{q_{j,k}}\right), -1\right]^T$. Consequently, $\mathcal{S}(-\beta_{j,2})$ is the probability that the participant correctly guesses the answer to the jth question when the participant does not possess all of the relevant skills required to answer the jth question (i.e., when $\prod_{k=1}^{K} \alpha_k^{q_{j,k}} = 0$). In addition, $1 - \mathcal{S}(\beta_{j,1} - \beta_{j,2})$ is the probability that the participant incorrectly answers

the jth question when the participant has all of the relevant skills (i.e., when $\prod_{k=1}^{K} \alpha_k^{q_{j,k}} = 1$).

Using the short-hand notation $p_{ij}(\boldsymbol{\beta}_j, \boldsymbol{\alpha}) = p\left(x_{ij} = 1 | \boldsymbol{\alpha}, \boldsymbol{\beta}_j, \mathbf{q}_j\right)$, it follows:

$$p(x_{ij}|\boldsymbol{\alpha}, \boldsymbol{\beta}_j, \mathbf{q}_j) = x_{ij} p_{ij}(\boldsymbol{\beta}_j, \boldsymbol{\alpha}) + (1 - x_{ij})(1 - p_{ij}(\boldsymbol{\beta}_j, \boldsymbol{\alpha})).$$

The conditional probability of all observed responses for the ith examinee, $\mathbf{x}_i$, given the skill attribute profile $\boldsymbol{\alpha}$ of the examinee is then given by the formula:

$$p(\mathbf{x}_i|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{j=1}^{J} p(x_{ij}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{q}_j).$$

2.1.3 Latent Skill Attribute Profile Probability Model

A Bernoulli latent skill attribute probability model (e.g., Maris, 1999) is assumed so that the probability that the kth latent skill, α_k, is present in the attribute pattern is given by the formula

$$p(\alpha_k|\eta_k) = \alpha_k \mathcal{S}(-\eta_k) + (1 - \alpha_k)(1 - \mathcal{S}(-\eta_k))$$

where η_k may be a free parameter. However, in this paper, η_k is assumed to be a constant. The probability, $p(\boldsymbol{\alpha})$, of a skill attribute profile, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]$, for an examinee is given by the formula: $p(\boldsymbol{\alpha}) = \prod_{k=1}^{K} p(\alpha_k|\eta_k)$.

2.1.4 Parameter Prior Model

The parameter prior for the two-dimensional parameter vector $\boldsymbol{\beta}_j$, associated with the jth question is a bivariate Gaussian density, $p(\beta_j)$, with two-dimensional mean vector $\boldsymbol{\mu}_j$ and two-dimensional covariance matrix $\sigma_\beta^2 \mathbf{I}_2$ for $j = 1, \ldots, J$ where $\mathbf{I}_2$ denotes the two-dimensional identity matrix. It is assumed that $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_J$ are known constants and σ_β^2 is a positive number. Let $\boldsymbol{\mu}_\beta = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_J]$. The joint distribution of $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_j]$ is specified by: $p(\boldsymbol{\beta}) = \prod_{j=1}^{J} p(\beta_j)$.

2.2 Parameter Estimation Method

The *complete-data likelihood function*, which assumes the latent skill attribute profile is observable, for the ith examinee is given by the formula:

$$p(\mathbf{x}_i, \boldsymbol{\alpha}|\boldsymbol{\beta}) = p(\boldsymbol{\alpha}) \prod_{j=1}^{J} p(x_{ij}|\boldsymbol{\alpha}, \boldsymbol{\beta}_j, \mathbf{q}_j). \tag{1}$$

However, since the latent skill attribute profile $\boldsymbol{\alpha}$ for an examinee is not observable we compute the *marginal likelihood function* for the ith examinee given by:

$$p(\mathbf{x}_i|\boldsymbol{\beta}) = \sum_{\boldsymbol{\alpha}} p(\mathbf{x}_i, \boldsymbol{\alpha}|\boldsymbol{\beta}) \tag{2}$$

where the summation is over all possible values of $\boldsymbol{\alpha}$ (i.e., every $\boldsymbol{\alpha}$ such that $p(\boldsymbol{\alpha}) > 0$).

The goal of the MAP estimation process is to find the parameter values $\boldsymbol{\beta}$ which maximize the posterior probability density function:

$$p(\boldsymbol{\beta}|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\beta})p(\boldsymbol{\beta})}{p(\mathbf{x}_1, \ldots, \mathbf{x}_n)}.$$

Towards this end, the MAP risk function, $\hat{\ell}_n(\boldsymbol{\beta})$, which is minimized to implement the MAP parameter estimation process is defined by the formula:

$$\hat{\ell}_n(\boldsymbol{\beta}) = -(1/n) \log \left(p(\boldsymbol{\beta})p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\beta})\right)$$

which since $p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\beta}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\beta})$ may be rewritten as:

$$\hat{\ell}_n(\boldsymbol{\beta}) = -(1/n) \log p(\boldsymbol{\beta}) - (1/n) \sum_{i=1}^{n} c(\mathbf{x}_i; \boldsymbol{\beta})$$

$$\text{where } \; c(\mathbf{x}_i; \boldsymbol{\beta}) = -\log \sum_{\boldsymbol{\alpha}} p(\mathbf{x}_i|\boldsymbol{\alpha}, \boldsymbol{\beta})p(\boldsymbol{\alpha}). \tag{3}$$

Because the calculation of the summation in (3) involves summing over different possible latent skill attribute patterns, the MAP risk function is not necessarily a unimodal function and may have multiple minimizers and saddlepoints (see Golden, 2020, 2022 for further discussion; also discussion near Eq. 5 of this paper).

To minimize $\hat{\ell}_n$, a gradient descent type method is used to search for critical points of $\hat{\ell}_n$ with the objective of finding a strict local minimizer of $\hat{\ell}_n$ that provides a good fit to the data generating process.

Define the *complete-data log likelihood* per observation, $\dot{c}(\mathbf{x}_i, \boldsymbol{\alpha}, \boldsymbol{\beta})$ by the formula: $\dot{c}(\mathbf{x}_i, \boldsymbol{\alpha}, \boldsymbol{\beta}) = -\log \left(p(\mathbf{x}_i|\boldsymbol{\alpha}, \boldsymbol{\beta})p(\boldsymbol{\alpha})\right)$. The *complete-data gradient* per observation, $\dot{\mathbf{g}}(\mathbf{x}_i, \boldsymbol{\alpha}, \boldsymbol{\beta})$, is given by the formula:

$$\dot{\mathbf{g}}(\mathbf{x}_i, \boldsymbol{\alpha}, \boldsymbol{\beta}) = -\sum_{j=1}^{J}(x_{ij} - p_{ij}(\boldsymbol{\beta}_j, \boldsymbol{\alpha}))\boldsymbol{\psi}(\boldsymbol{\alpha}, \mathbf{q}_j).$$

The gradient of $c(\mathbf{x}_i; \boldsymbol{\beta})$, $\mathbf{g}(\mathbf{x}_i, \boldsymbol{\beta})$, is given by the column vector-valued function:

$$\mathbf{g}(\mathbf{x}_i, \boldsymbol{\beta}) = \sum_{\boldsymbol{\alpha}} \dot{\mathbf{g}}(\mathbf{x}_i, \boldsymbol{\alpha}, \boldsymbol{\beta}) p(\boldsymbol{\alpha}|\mathbf{x}_i, \boldsymbol{\beta}), \quad p(\boldsymbol{\alpha}|\mathbf{x}_i, \boldsymbol{\beta}) = \frac{p(\mathbf{x}_i|\boldsymbol{\alpha}, \boldsymbol{\beta})p(\boldsymbol{\alpha})}{\sum_{\boldsymbol{\alpha}} p(\mathbf{x}_i|\boldsymbol{\alpha}, \boldsymbol{\beta})p(\boldsymbol{\alpha})}. \tag{4}$$

The summations over $\boldsymbol{\alpha}$ which are used in the computation of the gradient are often computationally intractable. However, the CDM used here has only five skills so the summations are straightforward since the summation only has $2^5 = 32$ terms.

To minimize the MAP risk function, the deterministic LBFGS algorithm (see Section 7.3.3 of Golden, 2020 for additional details) was used. The algorithm can be shown to converge to the set of critical points of $\hat{\ell}_n$ (e.g., Golden, 2020). Once a critical point is reached, then the Hessian of $\hat{\ell}_n$ can be evaluated at that point to check if the critical point is a strict local minimizer (e.g., Golden, 2020).

2.3 Asymptotic Statistical Theory for Confidence Intervals

In this section, we present explicit details regarding the calculation of confidence intervals. Here we assume that a parameter estimate $\hat{\boldsymbol{\beta}}_n$ has been obtained which is a strict local minimizer of $\hat{\ell}_n(\boldsymbol{\beta})$ in some (possibly very small) closed, bounded, and convex region of the parameter space Ω for all sufficiently large n. Furthermore, we assume that the expected value of $\hat{\ell}_n(\boldsymbol{\beta})$, $\ell(\boldsymbol{\beta})$, has a strict global minimizer, $\boldsymbol{\beta}^*$, in Ω. This setup of the problem thus allows for situations where ℓ has multiple minimizers, maximizers, and saddlepoints over the entire unrestricted parameter space. Given these assumptions, it can be shown (e.g., White, 1982; Golden, 2020, 2022) that $\hat{\boldsymbol{\beta}}_n$ in this case is a consistent estimator of $\boldsymbol{\beta}^*$.

Let $\mathbf{A}^*$ denote the Hessian of ℓ, $\mathbf{A}(\boldsymbol{\beta})$, evaluated at $\boldsymbol{\beta}^*$. Let $\mathbf{B}^*$ denote $\mathbf{B}(\boldsymbol{\beta}) = E\{\mathbf{g}(\tilde{\mathbf{x}}_i, \boldsymbol{\beta})(\mathbf{g}(\tilde{\mathbf{x}}_i, \boldsymbol{\beta}))^T\}$ evaluated at the point $\boldsymbol{\beta}^*$. It is well known (e.g., White, 1982; Golden, 2020, 2022) that if $\mathbf{A}^*$ and $\mathbf{B}^*$ are positive definite, then the asymptotic distribution of $\hat{\boldsymbol{\beta}}_n$ is a multivariate Gaussian with mean $\boldsymbol{\beta}^*$ and covariance matrix $(1/n)\mathbf{C}^* \equiv (1/n)(\mathbf{A}^*)^{-1}\mathbf{B}^*(\mathbf{A}^*)^{-1}$. In the special case where the model is correctly specified in the sense that the observed data is i.i.d. with common probability mass function $p(\mathbf{x}|\boldsymbol{\beta}^*)$, then the covariance matrix $\mathbf{C}^*$ may be computed using either $(\mathbf{A}^*)^{-1}$ or $(\mathbf{B}^*)^{-1}$.

It then follows that the *Hessian covariance matrix* $[\mathbf{A}^*]^{-1}$ is estimated by evaluating the Hessian of $\hat{\ell}_n$, $\bar{\mathbf{A}}_n(\boldsymbol{\beta})$, at $\hat{\boldsymbol{\beta}}_n$ where

$$\bar{\mathbf{A}}_n(\boldsymbol{\beta}) = \frac{1}{n\sigma_\beta^2}\boldsymbol{I}_q + (1/n)\sum_{i=1}^{n}\ddot{\mathbf{A}}_i(\boldsymbol{\beta}), \quad \ddot{\mathbf{A}}_i(\boldsymbol{\beta}) = \frac{d\sum_{\boldsymbol{\alpha}}\dot{\mathbf{g}}(\mathbf{x}_i,\boldsymbol{\alpha},\boldsymbol{\beta})p(\boldsymbol{\alpha}|\mathbf{x}_i,\boldsymbol{\beta})}{d\boldsymbol{\beta}} \tag{5}$$

Using the Missing Information Principle (Louis, 1982, see Section 13.2.5 of Golden, 2020 for a review) it can be shown that the right-most equation in (5) is the difference of two positive semidefinite matrices. Thus, it is possible that $\bar{\mathbf{A}}_n(\boldsymbol{\beta})$ has negative eigenvalues for a particular $\boldsymbol{\beta}$. However, if $n\sigma_\beta^2$ is sufficiently small ("high regularization case"), this can be used to ensure that $\bar{\mathbf{A}}_n(\boldsymbol{\beta})$ is positive definite for a particular $\boldsymbol{\beta}$ and particular sample size n.

The *OPG covariance matrix* $[\mathbf{B}^*]^{-1}$ is estimated by

$$\hat{\mathbf{B}}_n = (1/n)\sum_{i=1}^{n}\mathbf{g}(\tilde{\mathbf{x}}_i,\hat{\boldsymbol{\beta}}_n)(\mathbf{g}(\tilde{\mathbf{x}}_i,\hat{\boldsymbol{\beta}}_n))^T.$$

Unlike the Hessian covariance matrix case, adjusting $n\sigma_\beta^2$ does not directly effect the rank of the OPG covariance matrix unless the probability model is correctly specified.

The *Robust covariance matrix* $\mathbf{C}^*$ is estimated by $\hat{\mathbf{C}}_n = \hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n\hat{\mathbf{A}}_n^{-1}$. A 95% *Robust confidence interval* for the jth element of $\hat{\boldsymbol{\beta}}_n$, $\hat{\beta}_{n,j}$ has an estimated lower bound of $\hat{\beta}_{n,j} - 1.96\sqrt{\hat{C}_{n,j,j}/n}$ and an estimated upper bound of $\hat{\beta}_{n,j} + 1.96\sqrt{\hat{C}_{n,j,j}/n}$ where $\hat{C}_{n,j,j}$ is the jth on-diagonal element of $\hat{\mathbf{C}}_n$. Both Hessian and OPG confidence intervals can be computed in a similar manner by respectively replacing $\hat{\mathbf{C}}_n$ in these confidence interval formulas with $\hat{\mathbf{A}}_n^{-1}$ or $\hat{\mathbf{B}}_n^{-1}$.

3 Simulation Study

3.1 Methods

This study used an extraction of the Tatsuoka (1984) Fraction-Subtraction data set (George et al., 2016; Tatsuoka, 1984) consisting of 15 questions, 5 skills, and 536 students. The data set was fit using the previously described MAP estimation method to the previously described CDM.

The same multivariate Gaussian prior was used for all items. The mean of the Gaussian prior for an item was chosen such that both the guess and slip probabilities were equal to 0.354. The covariance matrix for the Gaussian prior was $\sigma_\beta^2\mathbf{I}$. For the high regularization case, $\sigma_\beta^2 = 900/n$ so that $1.96\sigma_\beta = 2.54$ corresponds to an informative prior which assigns high probability mass on a small region surrounding the Gaussian prior mean. For the low regularization case, $\sigma_\beta^2 = 9{,}000{,}000/n$ so that $1.96\sigma_\beta = 254$ corresponds to an uninformative prior which assigns high probability mass for a larger region surrounding the Gaussian prior mean. The

specific numerical choices for σ_β^2 were chosen based upon some preliminary pilot simulation studies.

To create the correctly specified model experimental condition, the simulation study involved sampling with replacement from a CDM fitted to the original data set to generate multiple bootstrap data sets, rather than sampling with replacement from the data set. Then, the CDM used to generate the data was fit to the bootstrap data sets to obtain parameter estimates for the correctly specified case. The **Q** matrix of the CDM used to generate the data was then modified by randomly flipping 20% of its elements and then this modified-CDM was fit to the bootstrap data sets to obtain parameter estimates for the misspecified case.

Final statistics were first computed in two ways. First, all of the bootstrap data samples ("all cases") were used. Second, only confidence intervals which could be reliably estimated ("identifiable cases") were used. The criterion for a reliably estimated confidence interval involved checking if: (1) parameter estimates used to estimate the confidence interval satisfied the numerical convergence criterion that the infinity norm of the gradient with respect to those parameter estimates was less than 0.00001, and (2) the covariance matrix used to estimate the confidence interval had a condition number less than 1000.

3.2 Results

For identifiable cases, fewer bootstrapped samples satisfied the local identifiability criterion for the low regularization case (60%), while nearly all satisfied the criterion for the high regularization case (99%), illustrating that situations where the covariance matrix is nearly singular were common in this study.

Figure 1 shows that bootstrap-averaged confidence intervals for item slip probabilities in the correctly specified case are wider for the low regularization case and narrower for the high regularization case. Table 1 provides a quantitative comparison of how estimated Type 1 error rates vary as a function of Hessian, Robust, and OPG estimation methods, presence of misspecification, degree of regularization, and bootstrap sample local identifiability inclusion criterion. For the low regularization (uninformative prior) case where the model is correctly specified and satisfies the local identifiability criterion, the estimated Type 1 error rate was close to the expected $p = 0.05$ level for all covariance matrix estimators, consistent with theory. For the high regularization (informative prior) case where the model is misspecified, the estimated Type 1 error rate was close to the $p = 0.05$ level only for the robust covariance matrix estimator, which was also consistent with theory. Finally, the OPG covariance matrix estimator for the low regularization case showed good performance in the presence of model misspecification. This latter result is not theoretically expected and we plan to pursue additional simulation studies to investigate its reliability.

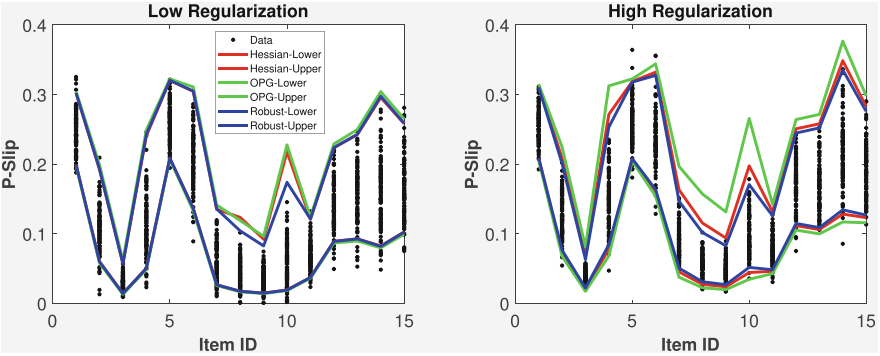


Fig. 1 Slip Probability Confidence Intervals: Correctly Specified Case. Averaged Hessian, OPG, and Robust 95% confidence intervals shown respectively using red lines, green lines, and blue lines are expected to contain approximately 95% of the individual bootstrap slip probability estimates which are denoted as black dots. Low regularization slip probability confidence intervals (left) are wider than high regularization slip probability confidence intervals (right)

Table 1 Estimated Type 1 error rate computed by counting the percentage of times bootstrap parameter estimates not included in an averaged 95% confidence interval. The numbers in parentheses are the number of bootstrap data sets, M, which satisfied a criterion specifying if the parameter estimates are identifiable. P-values close to 0.05 indicate agreement with asymptotic theory. Best performance was obtained for high regularization cases using the robust covariance matrix estimator

	All cases (M = 100)			Identifiable cases (M < 100)		
	Hessian	Robust	OPG	Hessian	Robust	OPG
Low Reg.						
Correct	0.066	0.072	0.062	**0.058** (57)	**0.057** (57)	**0.053** (57)
Misspecified	0.017	0.075	0.064	0.062 (63)	0.061 (63)	**0.054** (63)
High Reg.						
Correct	0.030	**0.049**	0.020	0.029 (99)	**0.048** (99)	0.019 (99)
Misspecified	0.038	**0.054**	0.023	0.039 (99)	**0.053** (99)	0.023 (99)

4 General Discussion

In this paper, we investigated how a regularization strategy influenced the quality of three different methods for estimating confidence intervals for MAP estimates in the presence of model misspecification. We found that increased levels of regularization tended to improve the quality of confidence interval estimation regardless of the presence of model misspecification when the Robust confidence interval estimation method was used. We also empirically showed that when the estimated covariance matrices are close to singular that the quality of the confidence interval estimators could be compromised.

These results suggest that regularization may be an important tool in practice for handling situations where near-singular covariance matrices are encountered,

and the Robust covariance matrix estimator may be important for ensuring reliable estimation of sampling error in such cases. In summary, these results emphasize the importance of examining the rank of the Hessian and OPG covariance matrix estimators to ensure reliable statistical inference. Results such as these will ultimately be important for developing robust CDM inference tools.

References

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The r package cdm for cognitive diagnosis models. *Journal of Statistical Software, 74*(2), 24. https://doi.org/10.18637/jss.v074.i02

Golden, R. M. (2020). *Statistical machine learning (texts in statistical sciences series)*. Chapman-Hall, CRC Press. https://www.routledge.com/Statistical-Machine-Learning-A-Unified-Framework/Golden/p/book/9781138484696

Golden, R. M. (2022). Estimating parameter confidence intervals in possibly misspecified parameter redundant models. https://mathpsych.org/presentation/769

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.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. https://projecteuclid.org/euclid.bsmsp/1200512988

Liu, Y., Xin, T., & Jiang, Y. (2021). Structural parameter standard error estimation method in diagnostic classification models: Estimation and application. *Multivariate Behavioral Research*, 1–20. https://doi.org/10.1080/00273171.2021.1919048

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 44*(2), 226–233. Retrieved October 1, 2022, from http://www.jstor.org/stable/2345828

Ma, W., & Jiang, Z. (2021). Estimating cognitive diagnosis models in small samples: Bayes modal estimation and monotonic constraints. *Applied Psychological Measurement, 45*(2), 95–111.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*(2), 187–212. https://doi.org/10.1007/BF02294535

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*(2), 177–195.

Paulsen, J., & Valdivia, D. S. (2022). Examining cognitive diagnostic modeling in classroom assessment conditions. *The Journal of Experimental Education, 90*(4), 916–933.

Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives, 16*(1), 1–17. https://doi.org/10.1080/15366367.2018.1435104

Tatsuoka, K. K. (1984). *Analysis of errors in fraction addition and subtraction problems*. Final report for NIE-G-81-0002.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50*(1), 1–25. https://doi.org/10.2307/1912526

Continuation Ratio Model for Polytomous Responses with Censored Like Latent Classes

Diego Carrasco, David Torres Irribarra, and Jorge González

Abstract Polytomous item responses are prevalent in background or context questionnaires of International large-scale assessments (ILSA). Responses to these types of instruments can vary in their symmetry or skewness. Zero inflation of responses can lead to biased estimates of item parameters in the response model and also to a downward bias in the conditional model when the zero inflated component is not accounted for in the model. In this paper, we propose to use a mixture continuation ratio response model to approximate the non-normality of the latent variable distribution. We use responses to bullying items from an ILSA study, which typically present positive asymmetry. The present model allows us to distinguish bullying victimization risk profiles among students, retrieve bullying victimization risk scores, and determine the population prevalence of the bullying events. This study also aims to illustrate how to fit a mixture continuation ratio model, including complex sampling design, thus expanding the modeling tools available for secondary users of large-scale assessment studies.

Keywords Continuation ratio model · Polytomous items · Item response theory · Bullying · Latent classes · Mixture models

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

D. Torres Irribarra
Escuela de Psicología, Pontificia Universidad Católica de Chile, Santiago, Chile

Millennium Nucleus on Intergenerational Mobility: From Modelling to Policy (MOVI), Santiago, Chile
e-mail: davidtorres@uc.cl

J. González
Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Santiago, Chile

Millennium Nucleus on Intergenerational Mobility: From Modelling to Policy (MOVI), Santiago, Chile
e-mail: jorge.gonzalez@mat.uc.cl

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_22

1 Introduction

Polytomous response distributions to multi-item instruments can vary in symmetry or skewness (Rhemtulla et al., 2012). In some cases, extreme responses can have a substantive interpretation. For example, the accumulation of the lowest response category of a list of symptoms can be interpreted as an absence of pathology. Similarly, an accumulation of responses in the lowest category of response to bullying victimization instruments can be expected from non-bullied students. In summary, there are cases where an extreme positive asymmetry in a response distribution across different items can be of interest.

However, common response models to polytomous items do not directly represent this accumulation to the lowest response category across all items. Within the international large-scale assessment studies (ILSA) (Rutkowski et al., 2010), the most popular response models to score responses to multi-items scales with ordered response categories are the partial credit model (PCM, Masters, 1982), the graded response model (GRM, Samejima, 1968) and confirmatory factor analyses (CFA, Jöreskog, 1969). None of these response models include a term to make interpretations and inferences to a high accumulation of responses to its lowest category. In the case of bullying instruments, a scenario where several items present a high accumulation of responses of this sort, previous studies have resorted to the creation of sums scores (e. g., Rutkowski et al., 2013). This latter strategy, mixed with the use of zero-inflated Poisson and negative binomial models, allows researchers to separate inferences regarding rates of bullying events and inferences regarding the absence of events for the accumulation of zeros in the sum score (Loeys et al., 2012).

Nevertheless, the sum score approach detaches the item side information of the response pattern. The sum score represents the person side of the response pattern to different items and the accumulation of zeros. Yet, the response prevalence among the items is lost, and the representation of measurement error is also lost. What can we do if we want to have all the information altogether? That is, to have the person side information regarding the absence of bullying, and rates of bullying, keeping-in the item side information of the relative prevalence among different bullying indicators. Moreover, what can we do, if we want to address these challenges, with ILSA studies? Essentially, we would need a response model that can give us what traditional IRT models do while adding an element to represent the zero responses, while also providing estimates generalizable to the surveyed population. Building from our previous work (Carrasco et al., 2022), in this paper, we propose a response model with mixtures to address the presented challenges with an applied example.

The paper is organized into six sections. We first describe the battery of items of our applied example and how responses to this instrument have been modeled in previous research (Sect. 2). In the following section (Sect. 3), we present our approach and its distinctive features to the selected application. In the Method section (Sect. 4), we describe the empirical data of our example, and we fit the presented model. We described the results in the Results section (Sect. 5). And

finally, in Sect. 6, we compare the presented model to other alternatives and point to future research.

2 Common Approaches to Model Responses to Bullying Victimization Items in Large-Scale Assessment

In the present work, we are interested in modeling responses to bullying items as a case where positive asymmetry is present, where a high frequency of zeros is interpretable. In particular, we have selected the "Students' experience of physical and verbal abuse at school" from the International Civic and Citizenship Education Study, abbreviated ICCS 2016 (Schulz et al., 2018). This battery of items consists of six bullying victimization events, to which students respond how often they have experienced these events in the last 3 months at their school. The response options are "not at all", "once", "2 to 4 times", and "5 times or more". Table 1 shows this battery of items and the observed percentage of response to each category for the pooled international sample of the study. In the literature on school bullying, these types of instruments are referred to as "bullying victimization measures" (Volk et al., 2014), and are used to represent the propensity of students to experience bullying events at their school. Other ILSA studies present similar batteries of bullying items, such as the "Student bullying scale" present in PIRLS 2016 (Martin et al., 2017), a similar variant "Student bullying scale" can be found in TIMSS 2019 (Yin &

Table 1 "Students' experience of physical and verbal abuse at school" in ICCS 2016

Frame	**During the last three months, how often did you experience the following situations at your school?** (Please tick only one box in each row.)					
		Not at all (%)	Once (%)	2–4 times (%)	5 times or more (%)	deltas
bul1	A student called you by an offensive nickname	45	26	15	14	−1.16
bul2	A student said things about you to make others laugh	44	27	18	12	−1.09
bul3	A student threatened to hurt you	81	11	5	3	0.30
bul4	You were physically attacked by another student	84	11	4	3	0.50
bul5	A student broke something belonging to you on purpose	80	15	4	2	0.53
bul6	A student posted offensive pictures or text about you on the Internet	90	7	2	1	0.90

Note: Response rates to each category, are obtained using the pooled international sample from ICCS 2016 (Schulz et al., 2018a, p. 236). We use "bul1-bul6" to refer to each specific item. "deltas" are delta dot parameters (Wu et al., 2016), reported in the technical report of the ICCS 2016 (Schulz et al., 2018, p. 164)

Fishbein, 2020), the "Exposure to bullying" is in PISA 2018 (OECD, 2019), and the "Violence within the school" can be found in ERCE 2019 (UNESCO, 2022). Thus, our selected case can shed light on common challenges across different ILSA studies regarding how to model bullying victimization responses.

Scale scores are derived from responses to the bullying scale in ICCS 2016 using a PCM. The generated logit scores are linearly transformed and release with a mean of 50 and a standard deviation of 10 units in the public data file (Schulz et al., 2018). Thus, when researchers are interested in studying relationships between the propensity to experience bullying with other factors, they do so using this generated scores available in the publicly released data from the study (e.g., Arroyo Resino et al., 2021; Schulz et al., 2018b; Tramontano et al., 2020). The application of the PCM also allows researchers to make inferences regarding the item side of the scale, and the prevalence of each bullying event in a multivariate way. By interpreting the delta parameter as a general location of item difficulty (Wu et al., 2016), one can make inferences regarding what is the riskier bullying event, the median item, and the most frequent bullying event among students. For instance, "bul6" is the riskier item from the battery. Students who have been bullied online are more likely to have also suffered from less risky bullying events (those items with lower delta estimates). In contrast, being teased (bul1, bul2) is a more common experience among students, and more than half of the students have experienced this event at least once.

A limitation of these generated scores is that they don't help researchers or secondary data users to represent the accumulation of responses to the lowest category across all items. Students who answer "not at all" to all items are the students who do not suffer from bullying at school. Using the derived scores, researchers can't easily separate the non-bullied students from those who suffer bullying in conditional models. The non-bullied students can be viewed as censored cases, representing students at the lowest level of bullying risk. Not accounting for these censored cases can lead to downward bias estimates for inferences models (e.g., conditioning bullying scores with other covariates) (Masyn, 2003). We believe researchers resort to using sum scores mixed with zero-inflated Poisson models, to surpass the previous limitation (e.g., Rutkowski et al., 2013; Rutkowski & Rutkowski, 2016). Assuming the items' responses fit closely with the PCM, the sum score can keep the order of bullying risk among respondents. Indeed, the PCM score over these items and its sum score are highly correlated ($r = .94$) within the ICCS 2016 study. Thus, the sum score provides an advantage for researchers interested in addressing inquiries regarding the person side of the responses. Using the sum score as the dependent variable, coupled with zero-inflated Poisson models, helps researchers to separate inferences between students at the lowest risk and students with some prevalence of bullying (Fu et al., 2013). Moreover, accounting for the "lowest risk" students in the conditional model helps to avoid the downward bias estimates in the inquiry of risk factors (Baetschmann & Winkelmann, 2017).

A limitation of the sum score approach under the presence of missing responses is that the simple sum score would classify students as "no risk," while the PCM scores could locate these cases differently, conditional on the prevalence of response to all items. Thus, a response model that can account for the "censored" side of the

distribution within the response model seems like an advantageous approach. In the next section, we describe a response model to address the present challenges.

3 Continuation Ratio Model with Mixtures to Account for Non-normality in the Latent Distribution

In this section we present an extension of a continuation ratio response model (CRM, Tutz, 2016), based on our earlier work (Carrasco et al., 2022). Continuation ratios are a way to formulate the logit link to the response categories and build a response model (de Boeck & Partchev, 2012). While partial credit models compare the log odds of adjacent categories and the graded response model relies on cumulative ratio logits, the proposed model uses the log odds of a category compared to all earlier categories for each item. This parametrization is referred to as a "decreasing order" continuation ratio (Gochyyev, 2015). To implement this response model, we use the expansion technique Gochyyev (2015) proposed, converting the original items with k responses into an expanded response matrix of *k-1* pseudo items in a wide data format (see Carrasco et al., 2022 for more details). In the present application θ_p represents the propensity of students to report being bullied, while δ_i represents how frequent the bullying event is across students, compared to earlier frequency options. If we code the response options of the bullying scale numerically as zero for "not at all," and 1, 2, 3 for "once," "2 to 4 times," and "5 times or more" respectively, the response model can be written as follows:

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

To expand the current model to account for the left censoring or non-normality of the latent distribution term, we include latent classes into the model. We include latent classes that preserve the item locations and divide the latent continuum into groups. For identification purposes, one latent mean is fixed to zero, while the rest of the latent means are freely estimated, while constraining the variance term of θ_p, $\zeta = 0$. This model specification is similar to a discrete survival model, where the frailty parameter is approximated using mixtures (Masyn, 2003, 2009), yet instead of modeling a "time-to-event" indicator expressing continuation ratios, our model is fitted into a series of continuation ratios used to represent the relative frequency of responses to bullying events. Figure 1 shows a schematic representation of the continuation ratio response model, and its' variant with mixtures using path diagrams. We will use the acronym M-CRM, to refer to this latter model.

In summary, the proposed model can be viewed as a non-parametric factor analysis, or as a case of a located latent class model (Masyn et al., 2010), with ordered category responses with continuation ratio logit links. This model

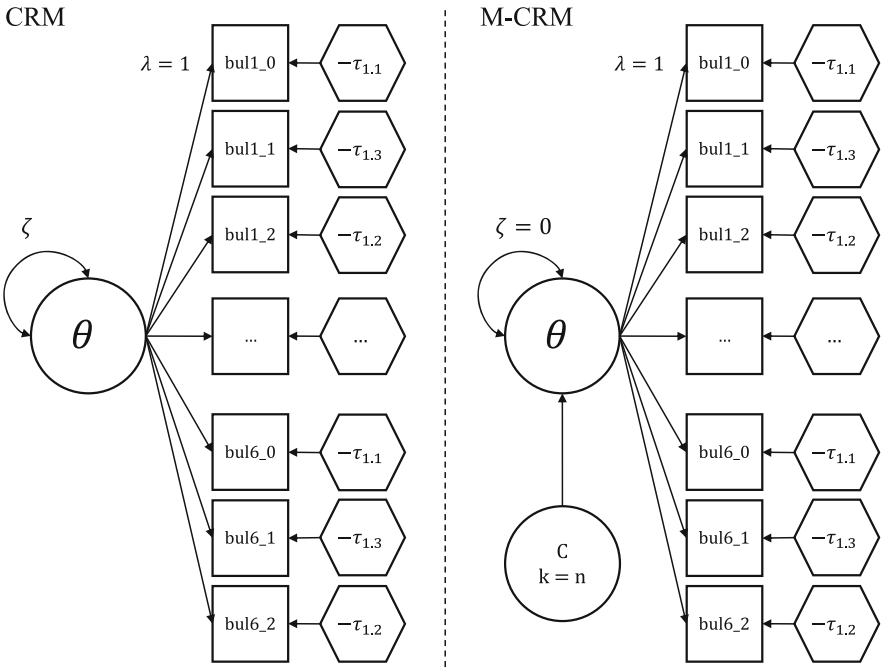


Fig. 1 Path diagrams of the continuation ratio response model (CRM), and its factor mixture variant (M-CRM). (**Note**: CRM is the continuation ratio response model presented in Eq. 1, and M-CRM is the path diagram of its factor mixture variant. θ represents the propensity of being bullied; bul1_0- bul1_2 are the pseudo items to represent the responses to item bul1 as continuation ratios; C represents the latent classes included in the model conditioning θ. ζ is the variance term for θ freely estimated in the CRM and constrained in M-CRM. λ represent the factor loadings in the model fixed to one in both models. $-\tau_{1.1}$ to $-\tau_{6.2}$ are the model thresholds that represent the item locations)

allows comparison of students on the θ continuum, and comparisons on the class membership. However, members of the same class are expected to have the same score on θ (Masyn et al., 2010). In the following section, we describe our illustration example.

4 Methods

4.1 Selected Data for Illustration

We use data from the International Civic and Citizenship Education Study from ICCS 2016. We are using responses to the "Students' experience of physical and verbal abuse at school" from ICCS 2016 (see Table 1). We are using the responses

from the Latin American participating samples, including Chile, Colombia, Dominican Republic, Mexico, and Perú. ICCS 2016 follows a two-stage sampling design, stratifying schools, and collecting responses from students from the same classroom within each selected school. This sample design reaches representative samples of 8th graders from each participating country, collecting responses of 3937–5609 students, and from 141 to 206 schools from each participating country (see Carrasco et al., 2022 for more details).

4.2 Analytical Strategy

To fit the proposed model, we use the pooled sample of selected countries and scaled the survey weights so each country sample contributes equally to the model estimates (Gonzalez, 2012). We generate the *k-1* pseudo items in wide format, converting the original responses to items bul1-bul6, into a continuation ratio dummy coded variables (bul1_0-bul6_2). These generated variables are our dependent variables in the response model. We use Taylor Series Linearization to account for the study sampling design, and pseudo maximum likelihood (Asparouhov, 2005) as implemented in Mplus 8.8 (Muthén & Muthén, 2017) to produce our estimates. Table 2 presents the Mplus code we used to fit the selected model.

To determine the number of latent classes in the model, we fit a series of models with 1–5 classes, and relied in the Vuong-Lo-Mendell-Rubin Likelihood Ratio test (VLMR-LRT) comparing k versus $k + 1$ classes (Nylund-Gibson & Choi, 2018). This later test favors the M-CRM with three classes over the M-CRM with four classes (VLMR-LRT $p = .02$). We described the results using the selected model.

5 Results

We describe the obtained results in three parts. First, the fixed effects estimates of the model are used to describe the relative frequency of the bullying events indicators. We analyze the person parameters approximated with mixtures in the second part to describe bullying risk profiles. Finally, we used the model expected proportions of the first pseudo item of each bullying indicator to describe the difference between the classes regarding their response profile. In Fig. 2, We use a modified version of an item-person map to summarize the main results.

5.1 Item Side

The results for the item side of the M-CRM are similar to those obtained with a continuation ratio response model with a continuous latent variable distribution

Table 2 Mplus syntax used to fit the CRM-C

```
TITLE:M-CRM-C3;

DATA:
FILE =
"bull_scale.dat";

VARIABLE:
NAMES =
id_i id_j id_s ws
bul1 bul2 bul3
bul4 bul5 bul6
bul1_2 bul1_1 bul1_0
bul2_2 bul2_1 bul2_0
bul3_2 bul3_1 bul3_0
bul4_2 bul4_1 bul4_0
bul5_2 bul5_1 bul5_0
bul6_2 bul6_1 bul6_0
;

MISSING=.;

CATEGORICAL =
bul1_2 bul1_1 bul1_0
bul2_2 bul2_1 bul2_0
bul3_2 bul3_1 bul3_0
bul4_2 bul4_1 bul4_0
bul5_2 bul5_1 bul5_0
bul6_2 bul6_1 bul6_0
;

USEVARIABLES =
bul1_2 bul1_1 bul1_0
bul2_2 bul2_1 bul2_0
bul3_2 bul3_1 bul3_0
bul4_2 bul4_1 bul4_0
bul5_2 bul5_1 bul5_0
bul6_2 bul6_1 bul6_0
;

IDVARIABLE = id_i;
WEIGHT = ws;
CLUSTER = id_j;
STRATIFICATION = id_s;
CLASSES = c(3);

TYPE = COMPLEX MIXTURE;
PROCESSORS = 4;
ESTIMATOR = MLR;
STARTS     = 100 10;
ALGORITHM = INTEGRATION;
LRTBOOTSTRAP = 100;

MODEL:
%overall%
!loadings
theta by bul1_0@1;
theta by bul1_1@1;
theta by bul1_2@1;
theta by bul2_0@1;
theta by bul2_1@1;
theta by bul2_2@1;
theta by bul3_0@1;
theta by bul3_1@1;
theta by bul3_2@1;
theta by bul4_0@1;
theta by bul4_1@1;
theta by bul4_2@1;
theta by bul5_0@1;
theta by bul5_1@1;
theta by bul5_2@1;
theta by bul6_0@1;
theta by bul6_1@1;
theta by bul6_2@1;

!variance;
theta@0;

!item locations
[bul1_0$1] (10);
[bul1_1$1] (11);
[bul1_2$1] (12);
[bul2_0$1] (20);
[bul2_1$1] (21);
[bul2_2$1] (22);
[bul3_0$1] (30);
[bul3_1$1] (31);
[bul3_2$1] (32);
[bul4_0$1] (40);
[bul4_1$1] (41);
[bul4_2$1] (42);
[bul5_0$1] (50);
[bul5_1$1] (51);
[bul5_2$1] (52);
[bul6_0$1] (60);
[bul6_1$1] (61);
[bul6_2$1] (62);

%c#1%
[theta*-1] (a1);

%c#2%
[theta@0] (a2);

%c#3%
[theta*1] (a3);

OUTPUT:
STANDARDIZED
CINTERVAL
RESIDUAL
TECH11;

SAVEDATA:
SAVE = FSCORES;
FILE =
crm_svy_3c_eap.dat;
```

Note: *id_i* is the unique student identifier; *id_j* is the unique school identifier; *id_s* is the unique strata identifier of the sampling, made distinct between countries, generating 375 unique pseudo strata; *ws* is the normalized survey weights vector, code preceded by "!" are comments

(see Carrasco et al., 2022), and to those obtained from the partial credit model (see Table 1). Accordingly, they lead to similar conclusions when identifying what are the most prevalent bullying event and what are the least prevalent bullying event across students. The item parameters are informative of the relative frequency of each bullying event across the pooled population of students from Chile, Colombia, Mexico, Dominican Republic and Perú. The most frequent bullying event among students is being mocked (bul2) and being called an offensive nickname (bul1). The most frequent and riskier bullying event is being shamed on the internet (bul6). Students who have suffered from being shame via offensive pictures and offensive texts on the internet are also more likely to have suffered from other bullying events, such as physical attacks and threats.

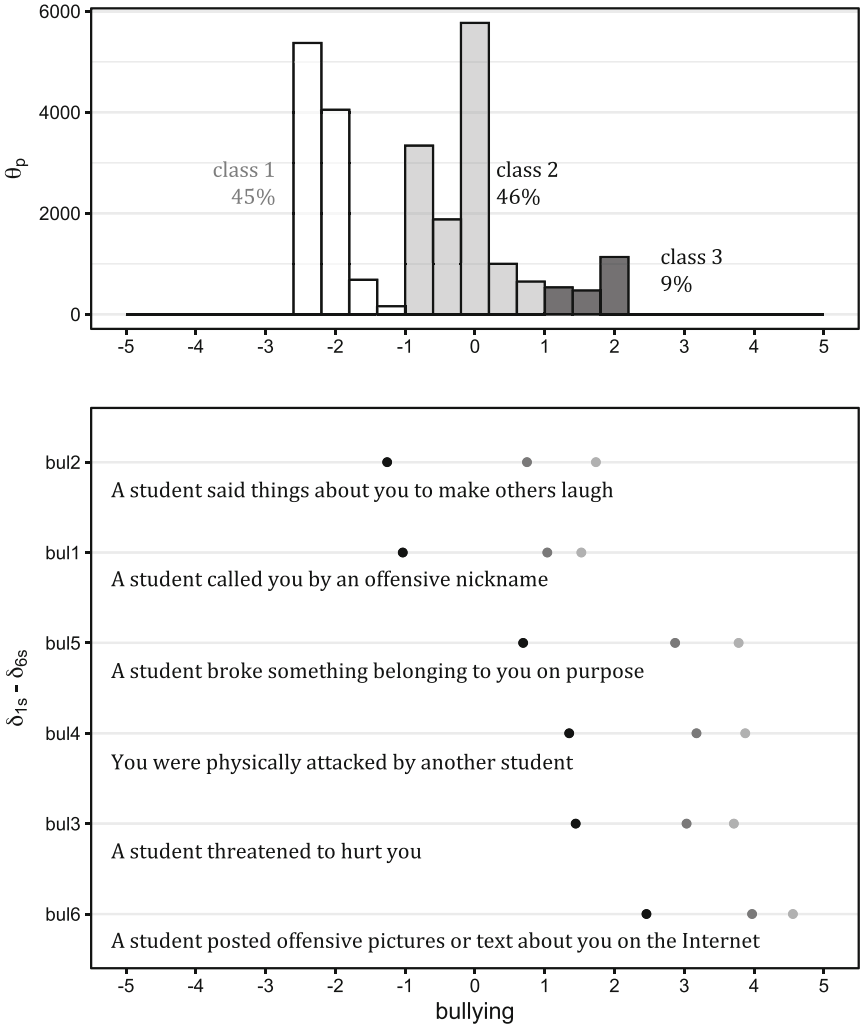


Fig. 2 Item person map of the bullying scale, using the estimates of the continuation response model with mixtures

5.2 Person Side

For illustration purposes, we fit a CRM, GRM, and PCM models to the same data, including the complex sampling design, estimating person and item parameters (θ, δ). The relative fit indices favor the PCM model ($BIC_{pcm} = 226970.254$; $BIC_{grm} = 227050.074$; $BIC_{crm} = 227746.389$, $BIC_{crm\text{-}c2} = 232008.366$, $BIC_{crm\text{-}c3} = 228503.806$). The EAP predictions are highly correlated between the PCM, GRM, CRM, and the M-CRM with three classes (r = .97–1.00), presenting

slightly smaller correlations to the M-CRM with two classes θ realizations (r = .89–.90). In general terms, the θ realizations of the selected model, the M-CRM with three classes, allow to order person in terms of their propensity to be bullied similar to the other response models.

A unique feature of the M-CRM with three classes model is the separation of the persons into mixtures. Using three latent classes, we can divide the propensity to be bullied into response profiles, a low-risk profile, a modal risk profile, and a high-risk profile, denominated class 1, class 2, and class 3 in Fig. 2, respectively. 45% of the students are in the lower risk profile, 46% in the modal risk profile, and 9% in the high-risk profile. In the next section, we illustrate what each of these profiles entails regarding their expected responses.

5.3 Expected Responses

In Table 3, we present the results of the expected responses to all the bullying event indicators, at their first response category, conditional to their latent class. We considered class 1 a low risk profile, class 2 a modal risk profile, and class 3 a high-risk bullying victimization profile. Students in the low risk are more likely to be mocked by their peers at school and they are very unlikely to be threatened, suffered from physical attacks, or shamed online. Students in the modal profile are more likely to be mocked but very unlikely to be shamed online and present some risk of threats and physical attacks. In contrast, students in the high-risk profile report to have suffered more frequent bullying events, and most of these students have experience threats and physical attacks. A distinctive feature of this profile is that these students are more likely to have experienced online shaming (bul6) than the rest of the other profiles.

Table 3 Expected proportions of "once" over "not at all" responses to bullying events

Bullying event	Item	Low risk	Modal risk	High risk
A student said things about you to make others laugh	bul2	0.25	0.78	0.96
A student called you by an offensive nickname	bul1	0.21	0.74	0.96
A student broke something belonging to you on purpose	bul5	0.05	0.33	0.79
You were physically attacked by another student	bul4	0.02	0.21	0.66
A student threatened to hurt you	bul3	0.02	0.19	0.64
A student posted offensive pictures or text about you on the Internet	bul6	0.01	0.08	0.39

Note: All standard errors were equal to or lower than .003

6 Conclusion and Discussion

Different response models are designed to account for the non-normal distribution of the person parameters and skewness across the responses to items (Reise et al., 2018). Among these, it is possible to differentiate between models changing the assumptions of the distribution of the person parameters or latent distributions and models including mixtures to account for the non-normality of the latent term. The present model is of the second kind.

We have illustrated how to fit a continuation ratio response model, including mixtures, with sampling design, to a set of bullying event indicators. These types of measures are commonly used in ILSA studies. The specified model accounts for the heterogeneity of students regarding their propensity to be bullied at school and uses mixtures to approximate the non-normal distribution of this heterogeneity. It helps to distinguish between low and high bullying victimization profiles while simultaneously providing information on the relative prevalence of each bullying event across students.

We believe the presented model has certain advantages. Unlike response models assuming normally distributed continuous latent variables, the mixtures present in the model can account for the floor effects (e.g., high accumulation of responses in the lowest categories) and the positive asymmetry across item responses. Similar models designed to account for zero inflation and asymmetry, such as the zero inflation graded response and the Davidian curve graded response (Smits et al., 2020; Wall et al., 2015), are competing options for the present case. However, these models were thought for scenarios where there is no presence of the attribute of interest (i.e., no pathology), or there is high asymmetry only. We believe the factor mixture variant we specified offers a less restrictive assumption regarding the prevalence of the frequency of the indicators in the least risky profile, allowing some prevalence of the bullying event indicators instead of no prevalence at all. The present model is similar to a located latent class Rasch model (Robitzsch & Steinfeld, 2018), that relies on continuation ratio logits instead of adjacent logits to model the responses. This model could be fitted using freely available software using the R library TAM (Kiefer et al., 2016), using the item expansion technique illustrated here. Thus, both applications should reach similar substantive results. Our presented approach has the advantage that it can be fitted using different latent variable modelling software, that allows users to include the complex sample survey of ILSA studies, and produce results generalizable to the population of students, present in the most popular ILSA studies.

Further research is needed to address if this model expansion has the same expected advantages in linking and equating of the CRM, as Kim (2016) suggested. Smits and colleagues (2020) assert that zero inflation leads to bias in slope and item locations of graded responses. The present factor mixture variant of the CRM could be another tool to address the high accumulation of responses on the left side of the distribution of bullying items.

Acknowledgments Research funded by the Fondo Nacional de Desarrollo Científico y Tecnológico FONDECYT N° 1201129 and FONDECYT N° 11180792; and Agencia Nacional de Investigacion y Desarrollo (ANID) Millennium Nucleus on Intergenerational Mobility: From Modelling to Policy (MOVI); (NCS2021072).

References

Arroyo Resino, D., Sandoval-Hernandez, A., & Eryilmaz, N. (2021). Characteristics of the schools resilient to violence. A study based on data from ICCS 2016 in Chile, Mexico and Colombia. *International Journal of Educational Research, 109*(July), 101839. https://doi.org/10.1016/j.ijer.2021.101839

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 12*(3), 411–434. https://doi.org/10.1207/s15328007sem1203_4

Baetschmann, G., & Winkelmann, R. (2017). A dynamic hurdle model for zero-inflated count data. *Communications in Statistics – Theory and Methods, 46*(14), 7174–7187. https://doi.org/10.1080/03610926.2016.1146766

Carrasco, D., Torres Irribarra, D., & González, J. (2022). Continuation ratio model for polytomous items under complex sampling design. In M. Wiberg, D. Molenaar, J. González, J.-S. Kim, & H. Hwang (Eds.), *Quantitative psychology*. Springer.

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software, 48*(Code Snippet 1). https://doi.org/10.18637/jss.v048.c01

Fu, Q., Land, K. C., & Lamb, V. L. (2013). Bullying victimization, socioeconomic status and behavioral characteristics of 12th graders in the United States, 1989 to 2009: Repetitive trends and persistent risk differentials. *Child Indicators Research, 6*(1), 1–21. https://doi.org/10.1007/s12187-012-9152-8

Gochyyev, P. (2015). *Essays in psychometrics and behavioral statistics*. University of California. https://escholarship.org/content/qt36b8p5nw/qt36b8p5nw.pdf

Gonzalez, E. J. (2012). Rescaling sampling weights and selecting mini-samples from large-scale assessment databases. *IERI Monograph Series Issues and Methodologies in Large-Scale Assessments, 5*, 115–134.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*(2), 183–202. https://doi.org/10.1007/BF02289343

Kiefer, T., Robitzsch, A., & Wu, M. (2016). *TAM: Test analysis modules*. https://cran.r-project.org/web/packages/TAM/TAM.pdf

Kim, S. (2016). Continuation ratio model in item response theory and selection of models for polytomous items. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang (Eds.), *Quantitative psychology research* (Vol. 196, pp. 1–13). Springer. https://doi.org/10.1007/978-3-319-38759-8_1

Loeys, T., Moerkerke, B., de Smet, O., & Buysse, A. (2012). The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. *British Journal of Mathematical and Statistical Psychology, 65*(1), 163–180. https://doi.org/10.1111/j.2044-8317.2011.02031.x

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2017). Methods and procedures in PIRLS 2016. In *Methods and procedures in PIRLS 2016*. https://timssandpirls.bc.edu/publications/pirls/2016-methods.html

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/BF02296272

Masyn, K. E. (2003). *Discrete-time survival mixture analysis for single and recurrent events using latent variables*. University of California Los Angeles. http://www.statmodel.com/download/masyndissertation.pdf

Masyn, K. E. (2009). Discrete-time survival factor mixture analysis for low-frequency recurrent event histories. *Research in Human Development, 6*(2–3), 165–194. https://doi.org/10.1080/15427600902911270

Masyn, K. E., Henderson, C. E., & Greenbaum, P. E. (2010). Exploring the latent structures of psychological constructs in social development using the dimensional-categorical spectrum. *Social Development, 19*(3), 470–493. https://doi.org/10.1111/j.1467-9507.2009.00573.x

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science, 4*(4), 440–461. https://doi.org/10.1037/tps0000176

OECD. (2019). *PISA 2018 results (volume III): What school life means for students' lives: Vol. III*. PISA, OECD Publishing. https://doi.org/10.1787/acd78851-en

Reise, S. P., Rodriguez, A., Spritzer, K. L., & Hays, R. D. (2018). Alternative approaches to addressing non-normal distributions in the application of IRT models to personality measures. *Journal of Personality Assessment, 100*(4), 363–374. https://doi.org/10.1080/00223891.2017.1381969

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373. https://doi.org/10.1037/a0029315

Robitzsch, A., & Steinfeld, J. (2018). Item response models for human ratings: Overview, estimation methods, and implementation in R. *Psychological Test and Assessment Modeling, 60*(1), 101–139. https://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2018_20180323/6_PTAM_IRMHR_Main__2018-03-13_1416.pdf

Rutkowski, L., & Rutkowski, D. (2016). *The relation between students' perceptions of instructional quality and bullying victimization* (In: Nilsen, T., Gustafsson, JE. (eds) teacher quality, instructional quality and student outcomes. IEA research for education, vol 2). Springer. https://doi.org/10.1007/978-3-319-41252-8_6

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher, 39*(2), 142–151. https://doi.org/10.3102/0013189X10363170

Rutkowski, D., Rutkowski, L., & Wild, J. (2013). Predictors of school violence internationally: The importance of immigrant status and other factors. In *5th IEA international research conference, 26–28 June 2013, Singapore*. http://www.iea.nl/fileadmin/user_upload/IRC/IRC_2013/Papers/IRC-2013_Rutkowski_etal.pdf

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *ETS Research Bulletin Series, 1968*(1), i–169. https://doi.org/10.1002/j.2333-8504.1968.tb00153.x

Schulz, W., Ainley, J., Fraillon, J., Losito, B., Agrusti, G., & Friedman, T. (2018a). *Becoming citizens in a changing world*. Springer. https://doi.org/10.1007/978-3-319-73963-2

Schulz, W., Ainley, J., Fraillon, J., Losito, B., Agrusti, G., & Friedman, T. (2018b). School contexts for civic and citizenship education. In *Becoming citizens in a changing world* (pp. 145–176). Springer. https://doi.org/10.1007/978-3-319-73963-2_6

Schulz, W., Carstens, R., Losito, B., & Fraillon, J. (2018). *ICCS 2016 technical report* (W. Schulz, R. Carstens, B. Losito, & J. Fraillon (Eds.)). International Association for the Evaluation of Educational Achievement (IEA).

Smits, N., Öğreden, O., Garnier-Villarreal, M., Terwee, C. B., & Chalmers, R. P. (2020). A study of alternative approaches to non-normal latent trait distributions in item response theory models used for health outcome measurement. *Statistical Methods in Medical Research, 29*(4), 1030–1048. https://doi.org/10.1177/0962280220907625

Tramontano, C., Nocentini, A., Palmerio, L., Losito, B., & Menesini, E. (2020). Mapping community, social, and economic risks to investigate the association with school violence and bullying in Italy. *Child Abuse and Neglect, 109*(April), 104746. https://doi.org/10.1016/j.chiabu.2020.104746

Tutz, G. (2016). Sequential models for ordered responses. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 1, pp. 139–151). CRC Press. https://doi.org/10.1201/9781315374512

UNESCO. (2022). *Manual de uso de las bases de datos Estudio Regional Comparativo y Explicativo (ERCE 2019)* (Issue Erce). https://unesdoc.unesco.org/ark:/48223/pf0000382518

Volk, A. A., Dane, A. V., & Marini, Z. A. (2014). What is bullying? A theoretical redefinition. *Developmental Review, 34*(4), 327–343. https://doi.org/10.1016/j.dr.2014.09.001

Wall, M. M., Park, J. Y., & Moustaki, I. (2015). IRT modeling in the presence of zero-inflation with application to psychiatric disorder severity. *Applied Psychological Measurement, 39*(8), 583–597. https://doi.org/10.1177/0146621615588184

Wu, M., Tam, H. P., & Jen, T.-H. (2016). Partial credit model. In *Educational measurement for applied researchers* (pp. 159–185). Springer Singapore. https://doi.org/10.1007/978-981-10-3302-5_9

Yin, L., & Fishbein, B. (2020). Creating and interpreting the TIMSS 2019 context questionnaire scales. In M. O. Martin, M. Von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report*. TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

A Three-Step Rectangular Latent Markov Modeling for Advising Students in Self-learning Platforms

R. Fabbricatore, R. Di Mari, Z. Bakk, M. de Rooij, and F. Palumbo

Abstract Recent years have seen a growing interest in using technology to provide adaptive learning environments. In this vein, (self)learning environments that offer an automatic recommendation system play a fundamental role in supporting students' learning with tailored feedback. In this aim, essential steps consist in collecting students' responses and diagnosing their learning state throughout the learning process. This contribution proposes a three-step rectangular Latent Markov modeling to assess students' abilities by analyzing sequences of response patterns to item-sets recorded at time intervals during the course. Each sequence corresponds to a measurement model that focuses on different topics. Furthermore, students' ability is conceived as a multivariate latent variable that refers to diverse skills. The proposed approach consists of a three-step procedure: carrying out a multivariate Latent Class IRT model at each time point to find homogeneous groups of students according to their ability level; computing the time-specific classification error probabilities; fitting weighted logistic regressions to investigate the effect of socio-demographic and psychological variables on the initial and transition probabilities using the entries of the inverse of the classification error matrices as weights (BCH correction).

R. Fabbricatore (✉)
Department of Social Sciences, University of Naples Federico II, Naples, Italy
e-mail: rosa.fabbricatore@unina.it

R. Di Mari
Department of Economics and Business, University of Catania, Catania, Italy
e-mail: roberto.dimari@unict.it

Z. Bakk · M. de Rooij
Department of Methodology and Statistics, Leiden University, Leiden, The Netherlands
e-mail: z.bakk@fsw.leidenuniv.nl; rooijm@fsw.leidenuniv.nl

F. Palumbo
Department of Political Sciences, University of Naples Federico II, Naples, Italy
e-mail: fpalumbo@unina.it

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_23

Keywords Latent variable models · Rectangular latent Markov modeling · Three-step approach · Learning statistics · Self-learning platforms

1 Introduction

Providing students with tailored feedback to support them during the learning process represents one of the main goals of students' ability assessment in education (Toomaneejinda, 2017). The development of proper recommendations and remediations is even more crucial when technology takes over the role of teachers, like in self-learning platforms. At this aim, the need for suitable statistical models to analyze the complex data structure that emerges from automatic student assessments stands out. Among them, latent variable models (Skrondal & Rabe-Hesketh, 2004) represent a relevant reference framework since students' ability can be conceived as a latent construct measured by a set of manifest indicators. Both parametric and non-parametric approaches were introduced in this framework (Bartolucci et al., 2015), allowing to address specific evaluation's aims.

Herein we rely on non-parametric approaches to identify latent classes; in this context, it means to find homogeneous groups of students according to their abilities. Non-parametric latent variable models are an ideal tool for developing accurate feedback during learning because they allow qualifying, in addition to quantifying, individual differences (McMullen & Hickendorff, 2018).

To be effective, recommendations should address all the aspects considered during an evaluation process, for example, topics, dimensions of students' ability (specific competencies), as well as any individual characteristic hypothesized to affect students' performance. Moreover, to understand intra-individual differences, it is also relevant to account students' abilities changing over time.

From a statistical modeling point of view, this means integrating multidimensionality (more latent variables defining students' ability), longitudinal design with a time-varying measurement model (different topics per time point), and the covariate effects on the students' progress in learning. The data consist of students' response patterns collected along the learning process and related to the specific competencies of interest, and a set of variables that refers to the individual students' characteristics.

In this vein, to include all the above-mentioned elements, the present contribution proposes a three-step rectangular latent Markov modeling. In particular, the three-step approach (Di Mari et al., 2016) allows managing different measurement models per time point. The novelty consists in adopting a rectangular formulation of the latent Markov model that enables different numbers of latent classes over time (Anderson et al., 2019).

The following section introduces the three-step rectangular latent Markov modeling, then an empirical application in the context of learning Statistics is presented in Sect. 3. Finally, some concluding remarks end the paper in Sect. 4.

2 Three-Step Rectangular Latent Markov Modeling

The proposed estimation procedure runs over the following three steps: a multidimensional latent class Item Response Theory (IRT) model allows detecting homogeneous groups of students according to their ability at each time point (Step 1); the time-specific class membership and classification error are computed (Step 2); a set of weighted logistic regressions are fitted to investigate the effect of the individual characteristics on the initial and transition probabilities (Step 3). The remainder of the section details the procedure's steps, whereas the R Code for the current application is provided in the Appendix.

2.1 Step 1: Multidimensional Latent Class IRT Model

The multidimensional latent class IRT (MLCIRT) models represent a semi-parametric extension of the traditional IRT models in that both the constraints of unidimensionality and continuous nature of the latent trait are released (Bartolucci, 2007). Given the matrix of students' response patterns, the MLCIRT model allows detecting sub-populations of homogeneous students according to their ability level, concurrently accounting for the multidimensional nature of students' ability and item characteristics (e.g., difficulty and discriminating power).

More formally, the vector $\boldsymbol{\Theta} = (\Theta_1, \Theta_2, \ldots, \Theta_D)'$ of the D latent variables, at each time $t = 1, \ldots, T$, follows a discrete distribution with $\boldsymbol{\xi}_1^{(t)}, \boldsymbol{\xi}_2^{(t)}, \ldots, \boldsymbol{\xi}_{k_t}^{(t)}$ vector of support points defining k_t latent classes, where k_t indicates the number of classes at the time t. For any t, $\boldsymbol{\pi}^{(t)} = \pi_1^{(t)}, \ldots, \pi_{k_t}^{(t)}$ are the prior probabilities of belonging to latent classes. Specifically, $\pi_c^{(t)} = P(\boldsymbol{\Theta}^{(t)} = \boldsymbol{\xi}_c^{(t)})$ with $c = 1, \ldots, k_t$, and $\sum_{c=1}^{k_t} \pi_c^{(t)} = 1$. At time $t = 2, ..., T$, the vector of class weights is obtained as $\boldsymbol{\pi}^{'(t)} = \boldsymbol{\pi}^{'(1)} \prod_{h=2}^{t} \boldsymbol{\Gamma}^h$, where $\boldsymbol{\Gamma}^t$ is the time-specific matrix of transition probabilities of order $k_{(h-1)} \times k_h$, given our working assumptions.

Without loss of generality and for the sake of space, we present the measurement part of the model only referring to the Generalized Partial Credit Model (GPCM; Bacci et al., 2014) among the IRT models. The used notation is typical of the IRT framework; not familiar readers can refer to Bartolucci et al. (2015) for more details. Thus, let us define $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_D^{(t)})$ as a realization of $\boldsymbol{\Theta}$ at time t, and $\theta_{cd}^{(t)}$ taking value in $\boldsymbol{\xi}_c^{(t)}$. The response $Y_{si}^{(t)}$ of the individual $s = 1, \ldots, n$ to a generic polytomous item i ($i \in \mathcal{I}$, the set of the items), with l_i response categories indexed from 0 to $l_i - 1$ and administered at time t, can be parameterized as follows according to the GPCM:

$$\mathrm{g}[P(Y_{si}^{(t)} = r|\boldsymbol{\Theta}^{(t)} = \boldsymbol{\xi}_c^{(t)})] = \log \frac{P(Y_{si}^{(t)} = r|\boldsymbol{\Theta}^{(t)} = \boldsymbol{\xi}_c^{(t)})}{P(Y_{si}^{(t)} = r-1|\boldsymbol{\Theta}^{(t)} = \boldsymbol{\xi}_c^{(t)})} = a_i \left(\sum_{d=1}^{D} \delta_{id} \theta_{cd}^{(t)} - b_{ir} \right),$$
$$r = 1, \ldots, l_i - 1;$$

where g(·) is the local logit link function; δ_{id} is a dummy variable equal to 1 if the item i measures the latent trait d; a_i and b_{ir} represent the discrimination and the item-step difficulty parameter, respectively.

Given k_t, the parameter estimation of multidimensional latent class IRT models is performed through the Expectation-Maximization algorithm (Dempster et al., 1977) using the R package `MultiLCIRT` (Bartolucci et al., 2014). It is worth noting that the choice of k_t is a ticklish issue and some possible solutions are discussed in Sect. 3.

2.2 Step 2: Modal Class Assignment and Classification Error

Since different measurement models are estimated for each time point, time-specific class membership and classification error probabilities are computed at this step. Following the modal assignment rule, each subject is assigned to the class corresponding to the highest posterior probability. More formally, the individual s is assigned to latent class g according to:

$$g^{(t)} = \underset{c=1,\ldots,k_t}{\text{argmax}}\ P(\boldsymbol{\Theta}_s^{(t)} = \boldsymbol{\xi}_c^{(t)} | \boldsymbol{Y}_s^{(t)} = \boldsymbol{y}_s^{(t)}),$$

where the posterior class probability can be expressed according to the Bayes's theorem as follows:

$$P(\boldsymbol{\Theta}_s^{(t)} = \boldsymbol{\xi}_c^{(t)} | \boldsymbol{Y}_s^{(t)} = \boldsymbol{y}_s^{(t)}) = \frac{\pi_c^{(t)} \prod_{d=1}^{D} \prod_{i \in I_d} P(Y_{si}^{(t)} = y_{si}^{(t)} | \Theta_{sd}^{(t)} = \theta_{cd}^{(t)})}{P(\boldsymbol{Y}_s^{(t)} = \boldsymbol{y}_s^{(t)})}.$$

For each time point, modal assignment estimates the predicted class $W_s^{(t)}$ allocating a weight $w_{sg}^{(t)} = P(W_s^{(t)} = g^{(t)} | \boldsymbol{Y}_s^{(t)} = \boldsymbol{y}_s^{(t)}) = 1$ and zero weight otherwise.

On the other hand, classification error can be evaluated through the conditional probability of the estimated class value conditional on the true one (Vermunt, 2010), resulting in the overall time-specific $\boldsymbol{D}^{(t)}$ matrix with elements:

$$P(W^{(t)} = c^{(t)} | \boldsymbol{\Theta}^{(t)} = \boldsymbol{\xi}_g^{(t)}) = \frac{\frac{1}{n_t} \sum_{s=1}^{n_t} P(\boldsymbol{\Theta}^{(t)} = \boldsymbol{\xi}_g^{(t)} | \boldsymbol{Y}_s^{(t)} = \boldsymbol{y}_s^{(t)}) w_{sg}^{(t)}}{P(\boldsymbol{\Theta}^{(t)} = \boldsymbol{\xi}_g^{(t)})},$$

where $c, g = 1, \ldots, k_t$ and n_t is the sample size at time t.

2.3 Step 3: BCH Correction to Account for Covariate Effect

In Step 3, according to the BCH correction (Bolck et al., 2004), the $\boldsymbol{D}$ matrices computed at the previous step are used in the weighted logistic regressions to find the effect of covariates on initial and transition probabilities.

More technically, using a multinomial logistic regression model, the probability of the estimated class membership $W^{(t)}$ at time t given the vector $\mathbf{Z}_s$ of P time-invariant individual covariates can be parameterized as follows:

$$P(W^{(t)} = c^{(t)}|\mathbf{Z}_s) = \frac{\exp(\gamma_{0c}^{(t)} + \sum_{p=1}^{P} \gamma_{pc}^{(t)} Z_{sp})}{\sum_{q=1}^{k_t} \exp(\gamma_{0q}^{(t)} + \sum_{p=1}^{P} \gamma_{pq}^{(t)} Z_{sp})}. \tag{1}$$

In order to model the probability $P(\boldsymbol{\Theta}^{(t)} = \boldsymbol{\xi}_g^{(t)}|\mathbf{Z}_s)$, for $t = 1, \ldots, T$, according to Bolck et al. (2004), we can express the probability $P(W^{(t)} = c^{(t)}|\mathbf{Z}_s)$ as a linear combination of $P(\boldsymbol{\Theta}^{(t)} = \boldsymbol{\xi}_g^{(t)}|\mathbf{Z}_s)$ considering the classification errors as weights:

$$P(W^{(t)} = c^{(t)}|\mathbf{Z}_s) = \sum_{g=1}^{k_t} P(\boldsymbol{\Theta}^{(t)} = \boldsymbol{\xi}_g^{(t)}|\mathbf{Z}_s) P(W^{(t)} = c^{(t)}|\boldsymbol{\Theta}^{(t)} = \boldsymbol{\xi}_g^{(t)}). \tag{2}$$

Let be $e_{sc}^{(t)} = P(W^{(t)} = c^{(t)}|\mathbf{Z}_s)$, $a_{sg}^{(t)} = P(\boldsymbol{\Theta}^{(t)} = \boldsymbol{\xi}_g^{(t)}|\mathbf{Z}_s)$, and $d_{gc}^{(t)} = P(W^{(t)} = c^{(t)}|\boldsymbol{\Theta}^{(t)} = \boldsymbol{\xi}_g^{(t)})$ element of matrices $\boldsymbol{E}^{(t)}$, $\boldsymbol{A}^{(t)}$, and $\boldsymbol{D}^{(t)}$, respectively. The matrix notation of Eq. 2 is:

$$\boldsymbol{E}^{(t)} = \boldsymbol{A}^{(t)} \boldsymbol{D}^{(t)}.$$

Accordingly, we can obtain the matrix $\boldsymbol{A}^{(t)}$ with the probabilities of true class membership given individual covariates as follows:

$$\boldsymbol{A}^{(t)} = \boldsymbol{E}^{(t)} \boldsymbol{D}^{(t)^{-1}}$$

Thus, using the entries of the inverse of the $\boldsymbol{D}^{(t)}$ matrix as observation weights during the estimation of the multinomial regression in Eq. 1, we can obtain regression parameters referring to the probability $P(\boldsymbol{\Theta}^{(t)} = \boldsymbol{\xi}_g^{(t)}|\mathbf{Z}_s)$ (Vermunt, 2010).

Specifically, in the proposed approach, a multinomial regression with time 1 classification as dependent variable allows evaluating the covariate effect on initial probability. On the other hand, to estimate the effect of covariates on all the possible transitions over time, $\sum_{t=1}^{T-1} k_t$ multinomial regressions are required, each considering the individuals belonging to one of the latent classes emerged at time t as sample (in total k_t sub-samples for each time point) and the corresponding classification at time $t + 1$ as dependent variable.

3 Application: Learning Statistics in Non-STEM Degree Programs

In what follows, we present an application of the proposed approach in the context of Learning Statistics in non-STEM[1] degree programs. The study involved $N = 202$ Italian students (83.6% female; age: mean = 19.7, sd = 2.77) enrolled in the first year of the psychology course at the University of Naples Federico II, attending the introductory Statistics course.

Data collection was carried out via the Moodle platform and consisted of three waves, each focusing on different statistical topics: descriptive statistics, graphs, tables, and Gaussian distribution at Time 1, probability and random variables at Time 2, hypothesis testing and bivariate statistics at Time 3. Students' ability was conceived as a multidimensional latent variable according to three Dublin descriptors (Gudeva et al., 2012): understanding of theoretical concepts (Knowledge), ability to apply the knowledge to solve exercises (Application), and critical skills (Judgment). For each wave, students were asked to respond to 30 multiple-choice questions, 10 for each considered Dublin descriptor, which had four answer options and three different response scores: totally correct answers received two credits, partially correct answers received one credit, wrong answers received no credit. Blank responses were treated as missing values.

The analyzed data also comprises cognitive and psychological variables affecting learning Statistics, which were assessed by means of psychometric scales at the beginning of the course. It is worth noting that psychological variables assume a fundamental role during learning, affecting students' performance and achievement. Several studies (see, for example, Chiesi & Primi, 2010) highlighted the relevance of some of these factors especially for subjects perceived as frightening like Statistics for students enrolled in non-STEM degree courses. According to the existing literature, we accounted for the effect of math knowledge, statistical anxiety, attitudes toward Statistics, self-efficacy, and engagement on initial and transition probabilities. A more in-depth description of the data at hand can be found in Fabbricatore et al. (2022).

3.1 Results

In this section, we describe the results obtained by applying the proposed approach in the context of learning Statistics.

Firstly, we ensured scale comparability across time in order to compare the classifications obtained in the different time points. To this aim, we carried out an IRT factor analysis on the whole set of items to assess item characteristics

[1] STEM is the abbreviation for Science, Technology, Engineering, and Mathematics.

Table 1 IRT factor analysis results: Fit statistics for nested models to test parallel item similarity

		BIC	χ^2	df	p-value
Knowledge	Constrained	6411.63			
	Unconstrained	6436.25	4.30	6	0.64
Application	Constrained	6891.54			
	Unconstrained	6919.74	0.73	6	0.99
Judgment	Constrained	6735.15			
	Unconstrained	6760.20	3.88	6	0.69
Multidimensional	Constrained	20041.1			
	Unconstrained	20116.4	11.47	18	0.87

Note: Constrained = equal parameters across reference (parallel) items

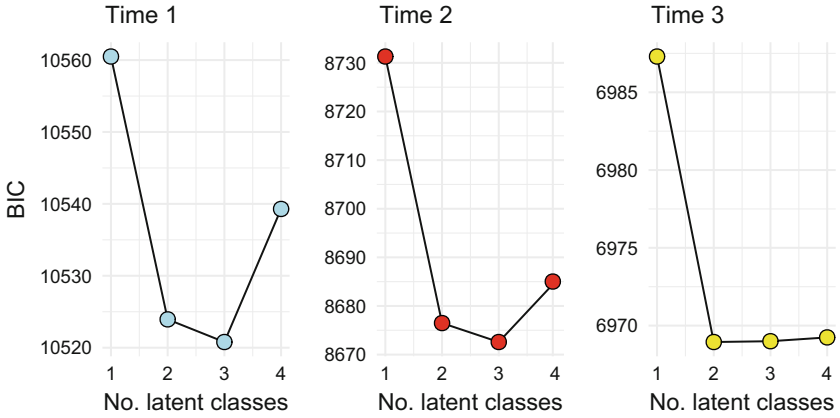


Fig. 1 BIC values for different number of latent classes at each time point

and select, for each dimension, the three items (one per time point) with the most similar characteristics in terms of difficulty and discrimination parameters. These "parallel items" can be used as reference items for identifiability issues in the multidimensional latent class IRT models at Step 1 to guarantee scale comparability over time. We also tested their similarity through a χ^2 test comparing nested IRT factor models where the constrained one is that with imposed equal parameters across the reference (parallel) items. Table 1 shows the results for both unidimensional and multidimensional models, all pointing at a not significant difference between the nested models and thus moving in favor of scale comparability.

Given the above results, we considered the parallel items as the reference for model identifiability in the multidimensional latent class IRT models in Step 1.

The number of latent classes for each time point can be defined according to theoretical assumptions or a data-driven approach. Herein, we employed the Bayesian Information Criterion (BIC), pointing at three latent classes at Time 1 and Time 2 and two latent classes at Time 3 (see Fig. 1).

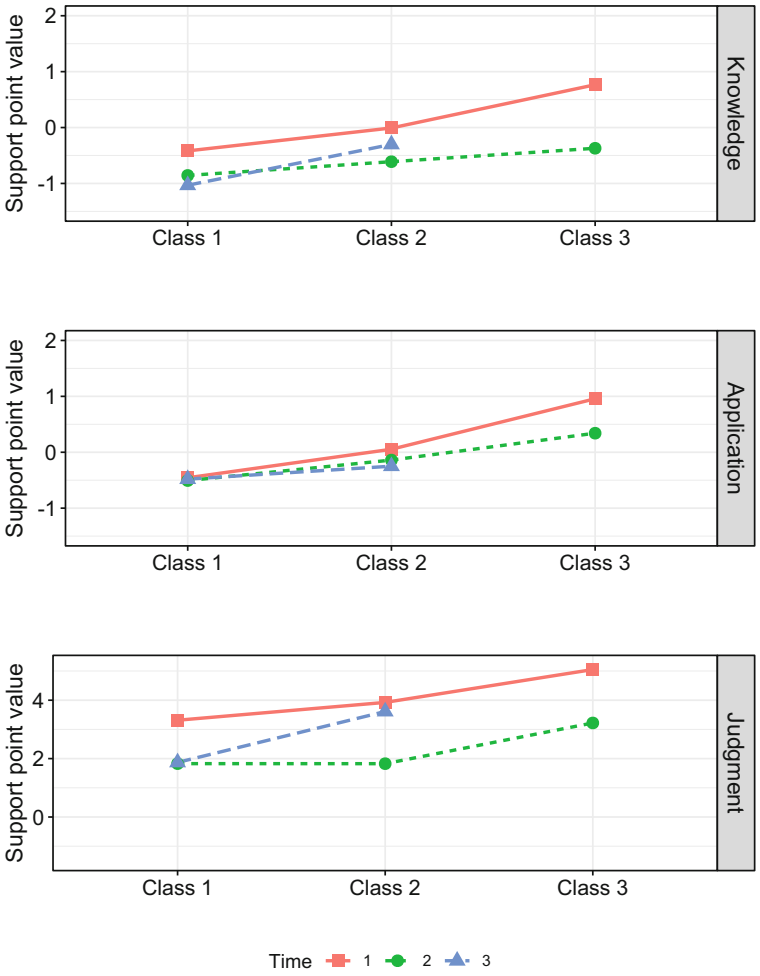


Fig. 2 Class profiles for the selected models. Support point value on the y-axis indicates the level of ability of students belonging to the considered latent classes. Accordingly, Class 1, Class 2, and Class 3 indicate low-, medium-, and high-ability learners, respectively

Looking at class profiles in Fig. 2, corresponding to the best models in terms of BIC, we can see that latent classes are increasingly ordered according to all the latent trait dimensions at each time point. Hence, Class 1, Class 2, and Class 3 indicate low, medium, and high levels of ability, respectively. Moreover, scale comparability allows affirming that students' ability was higher in Time 1 (descriptive statistics) than in Time 2 and Time 3, especially in Knowledge and Judgment. In contrast, a smaller difference in ability levels over time was found for Application. Note also

that the level of ability associated to Class 2 at Time 3 is very similar to the ability level of Class 2 at Time 1.

Regarding classification error probabilities computed in Step 2, the following $\boldsymbol{D}$ matrices resulted:

$$\boldsymbol{D}^{(1)} = \begin{bmatrix} 0.834 & 0.165 & 0.000 \\ 0.071 & 0.890 & 0.039 \\ 0.000 & 0.181 & 0.819 \end{bmatrix}; \ \boldsymbol{D}^{(2)} = \begin{bmatrix} 0.819 & 0.176 & 0.005 \\ 0.049 & 0.845 & 0.106 \\ 0.002 & 0.049 & 0.950 \end{bmatrix};$$

$$\boldsymbol{D}^{(3)} = \begin{bmatrix} 0.977 & 0.023 \\ 0.031 & 0.969 \end{bmatrix}$$

with the elements on the main diagonals providing evidence for an accurate classification at each time point.

The inverse of the obtained $\boldsymbol{D}$ matrices was used for the estimation of covariate effects in Step 3. Results showed that sex, math knowledge, and engagement significantly affect initial classification probabilities, whereas no significant effects were found for statistical anxiety, attitudes toward Statistics, and self-efficacy. In particular, females had a lower probability of being in class 2 ($\gamma_2 = -1.44$, p-value $= 0.058$) and Class 3 ($\gamma_3 = -2.77$, p-value $= 0.001$) with respect to Class 1 than males, highlighting an impairment in the level of ability according to sex at Time 1. Moreover, higher level of math knowledge was associated with a greater probability to be in Class 2 ($\gamma_2 = 0.08$, p-value $= 0.03$) and 3 ($\gamma_3 = 0.31$, p-value < 0.001), and thus a medium and high performance in Statistics. Also students' engagement in Statistics positively affected students performance, increasing the probability to be in Class 2 ($\gamma_2 = 0.73$, p-value $= 0.02$) rather than in Class 1.

The effect of covariates on transition probabilities was depicted in Fig. 3. Note that only the significant effects were reported, using red color for negative effects and green color for positive ones. Moreover, because some students dropped out during learning, we considered an additional class (Dropped) for Time 2 and Time 3 at this step.

Looking at transitions from Time 1 to Time 2, we can see that a lower level of engagement increased the risk of dropout for students belonging to Class 1, who have low ability levels. Moreover, math knowledge positively affected ability change over time, fostering the transitions of students from Class 2 and Class 3 at Time 1 to Class 3 at Time 2, namely the class with the highest level of ability.

In addition, Statistical anxiety and attitudes toward Statistics, although not significantly affecting initial classification probabilities, revealed to have a significant effect on the transitions. Specifically, feeling anxious during a Statistics test and considering Statistics a difficult subject reduced the probability of students belonging to Class 2 and Class 3 at Time 1 to move in Class 2 and Class 3 rather than in Class 1 at Time 2, thus negatively affected their performance. On the other hand, positive feelings concerning Statistics (affective attitudes) increased the probability of students in Class 2 at Time 1 to move in Class 2 rather than Class 1 at Time 2 (positive effect on performance).

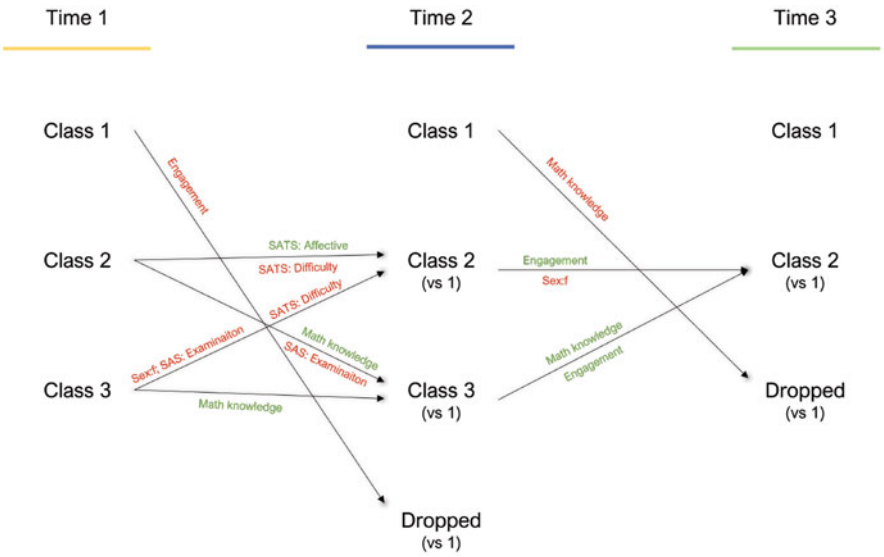


Fig. 3 Significant covariate effects on transitions. Note that Class 1 always represents the reference class. Negative effects are depicted in red, whereas positive effects are in green; SATS = Attitudes toward Statistics; SAS = Statistical anxiety

Conversely, transitions from Time 2 to Time 3 were affected only by math knowledge and engagement among the considered cognitive and psychological variables. We can speculate that this result could be related to the difficulty of the topics at Time 3, particularly requiring basic ability in math and students' effort in studying Statistics to perform better. Regarding the sex variable, we found that it also affected some transition probabilities, in addition to the initial ones, again underlying impairment in favor of males.

4 Conclusion

The proposed three-step rectangular latent Markov modeling represents a valuable statistical tool for analyzing student ability assessment data to develop tailored feedback in self-learning platforms. Indeed, it addresses several issues that are typical of ability assessment on self-learning platforms, which is based on a different measurement model per time point, different item characteristics (e.g., item difficulty and discrimination), and multiple ability dimensions.

Moreover, the proposed approach allows combining cross-sectional and longitudinal information, identifying students' strengths and weaknesses in comparison with their peers for each topic (cross-sectional) and understanding students' progress over time (longitudinal). In this regard, the rectangular formulation of

Latent Markov models also accounts for changes leading to different nature and number of latent classes or, as in our application, the presence of dropouts.

Therefore, model parameters can be used to provide students with adaptive feedback at different levels: according to the ability dimensions, the topics, peer performance, and progress over time. Also individual characteristics such as demographic, psychological, and cognitive factors affecting learning can be integrated into the model, allowing to develop motivational feedback along with formative one.

In this vein, the application in the context of learning Statistics shed light on the amount of useful information provided by the model with the aim of encouraging researchers to employ such a model when dealing with complex evaluations of students' ability.

Current developments work on a Maximum Likelihood (ML) correction for the third step of our approach, where a rectangular latent Markov model with class assignments as a single indicator and known error probabilities is used to estimate the structural part of the model.

Appendix

R Code for the current application.

Step 1: Multidimensional latent class IRT model

```
1
2  library(readxl)
3  library(MultiLCIRT)
4  library(dplyr)
5  library(mclust)
6  library(mclogit)
7
8  # Read the data file
9  Data = read_excel("data_matrix.xlsx", na = "999")
10
11 # Select items of Time 1 and order them according to the
       considered dimensions
12 Data_lc_r = dplyr::select(Data, starts_with('T1_S_'))
13 K = dplyr::select(Data_lc_r, ends_with('_K'))
14 A = dplyr::select(Data_lc_r, ends_with('_A'))
15 J = dplyr::select(Data_lc_r, ends_with('_J'))
16 Data_lc_r = cbind(K, A, J)
17
18 # Define the matrix with item indices according to the measured
       dimensions (for each dimension, the first item is the
       reference for model identifiability)
19 multi = rbind(c(3, 1, 2, rep(4:10)), c(14, rep(11:13),
20         rep(15:20)), c(23, 21, 22, rep(24:30)))
21
22 # Model selection (compare models with different number of
       classes) following the GPCM
```

```
GPCM = list()
for (i in 1:5) {
  GPCM[[i]] <- est_multi_poly(Data_lc_r, k = i, link = 2 ,
             disc = 1, difl = 0, output = T, multi = multi)
}

BIC_value = c(GPCM[[1]]$bic, GPCM[[2]]$bic, GPCM[[3]]$bic,
             GPCM[[4]]$bic, GPCM[[5]]$bic)

# Best model according to the BIC
GPCM3 = GPCM[[3]]

# Model parameters
GPCM3$piv # Class weights
GPCM3$Th # Matrix of support points
GPCM3$Bec # Item difficulty parameters
GPCM3$gac # Item discriminating parameters
GPCM3$Pp # Matrix of posterior probabilities

# Repeat lines 11-40 for Time 2 and Time 3 and obtain: "
    Data_lc_r2" and "Data_lc_r3" (datasets); "GPCM3_t2" and "
    GPCM2_t3" (MultiLCIRT model output)
```

Step 2: Modal class assignment and classification error

```
# Classify the observations according to the posterior class
    probabilities
Data_lc_r$Id = rep(1:nrow(Data_lc_r))
Data_lc_r$Clus1 = data.frame(rep(0, nrow(Data_lc_r)))
for (j in 1:nrow(Data_lc_r)) {
  Data_lc_r$Clus1[j,] = which.max(GPCM3$Pp[j,])
}

# Repeat lines 43-48 for Time 2 and Time 3 and obtain: "
    Data_lc_r2$Clus2" and "Data_lc_r3$Clus3"

# Create a matrix n x T with class assignments
total_class = as.data.frame(left_join(Data_lc_r[, c("Id",
    "Clus1")], Data_lc_r2[, c("Id", "Clus2")], by = c("Id")))
total_class = as.matrix(left_join(total_class[, c("Id", "Clus1",
    "Clus2")], Data_lc_r3[, c("Id", "Clus3")], by = c("Id")))

# Create a function for the modal D matrix computation
Dmatrix = function(outmodel){

classweight = outmodel$piv
numobservation = nrow(outmodel$Pp)
numofclasses = length(outmodel$piv)
posteriorprob = as.matrix(outmodel$Pp)

W = posteriorprob == outer(apply(posteriorprob,1, max),
    rep(1,numofclasses))
num = (t(posteriorprob) %*% W)/numobservation
Dmatrix = num/classweight

```

```
return(Dmatrix)
}

# Calculate the D matrix for each Time and store the results in
    the cD array
cD = array(NA, c(3,3,3))
    cD[,,1] = Dmatrix(GPCM3)
    cD[,,2] = Dmatrix(GPCM3_t2)
    cD[1:2,1:2,3] = Dmatrix(GPCM2_t3)
```

Step 3: BCH correction to account for covariate effect

```
# Cross-tabulate class assignments from previous steps to obtain
    initial and transitions as if the assignments were
    realizations of an observed Markov chain
inistart = table(total_class[,1+1], useNA = "always")/sum(table(
    total_class[,1+1], useNA = "always"))
PI2 = table(total_class[,1+1],total_class[,1+2], useNA = "always"
    )/rowSums(table(total_class[,1+1],total_class[,1+2], useNA =
    "always"))
PI3 = table(total_class[,1+2],total_class[,1+3], useNA = "always"
    )/rowSums(table(total_class[,1+2],total_class[,1+3], useNA =
    "always")) # Dropout class for NAs at Time 2 and Time 3
total_class_recod = total_class[,-1]
N = dim(total_class_recod)[1]
for(t in 1:3){
  total_class_recod[is.na(total_class_recod[,t]),t] = max(
    total_class_recod[,t],na.rm=T)+1
}

modal_class1 = mclust::unmap(total_class_recod[,1])
modal_class2 = mclust::unmap(total_class_recod[,2])
modal_class3 = mclust::unmap(total_class_recod[,3])

# Create "individual" transitions
iK = c(3,3,2) # Number of latent classes for each time point
PI2_dep = array(0,c(N,(iK[2]+1),iK[1]))
for(n in 1:N){
  PI2_dep[n,,] = t((modal_class1[n,])%*%t(modal_class2[n,]))
}

PI3_dep = array(0,c(N,(iK[3]+1),iK[2]+1))
for(n in 1:N){
  PI3_dep[n,,] = t((modal_class2[n,])%*%t(modal_class3[n,]))
}

# Create BCH weights from classification error probabilities
cDexp2 = diag(4)
cDexp2[1:3,1:3] = cD[,,2]
cDexp3 = diag(3)
cDexp3[1:2,1:2] = cD[1:2,1:2,3]

wei1 = diag(solve(cD[,,1]))[total_class_recod[,1]]
wei2 = diag(solve(cDexp2))[total_class_recod[,2]]
wei3 = diag(solve(cDexp3))[total_class_recod[,3]]
```

```
116
117 # Select the covariates from the dataset
118 covar = dplyr::select(Data, c("Sex", "PMP_Total",
119     "SAS_Examination", "SAS_Interpretation", "SATS_Affect",
120     "SATS_Difficulty", "MSLQ_SelfEfficacy", "ENG"))
121
122 # Estimate covariate effect at Time 1
123 df_t1 = data.frame(y = factor(mclust::map(modal_class1)), covar)
124 out_t1 = mblogit(y ~ ., data = df_t1, weights = wei1)
125
126 # Estimate covariate effect on transitions at Time 2
127 # Starting in state 1 (first row of transition matrix), first (
      arrival) state as reference
128 df_t2_s1 = data.frame(y = factor(mclust::map(PI2_dep[,,1])),
      covar)
129 out_t2_s1 = mblogit(y ~ ., data = df_t2_s1, weights = wei2)
130
131 # Repeat lines 128-129 for each sub-sample of individuals defined
       by the latent classes emerged at Time t considering the
      corresponding classification at Time t+1 as dependent
      variables to estimate the covariate effect on all the other
      transitions at Time 2 and Time 3
```

References

Anderson, G., Farcomeni, A., Pittau, M. G., & Zelli, R. (2019). Rectangular latent markov models for time-specific clustering, with an analysis of the wellbeing of nations. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 68*(3), 603–621.

Bacci, S., Bartolucci, F., & Gnaldi, M. (2014). A class of multidimensional latent class IRT models for ordinal polytomous item responses. *Communications in Statistics-Theory and Methods, 43*(4), 787–800.

Bartolucci, F. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika, 72*(2), 141–157.

Bartolucci, F., Bacci, S., & Gnaldi, M. (2014). MultiLCIRT: An R package for multidimensional latent class item response models. *Computational Statistics & Data Analysis, 71*, 971–985.

Bartolucci, F., Bacci, S., & Gnaldi, M. (2015). *Statistical analysis of questionnaires: A unified approach based on R and stata* (Vol. 34). CRC Press.

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis, 12*(1), 3–27.

Chiesi, F., & Primi, C. (2010). Cognitive and non-cognitive factors related to students' statistics achievement. *Statistics Education Research Journal, 9*(1), 6–26.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–22.

Di Mari, R., Oberski, D. L., & Vermunt, J. K. (2016). Bias-adjusted three-step latent Markov modeling with covariates. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(5), 649–660.

Fabbricatore, R., Bakk, Z., Di Mari, R., de Rooij, M., & Palumbo, F. (2022). Students' proficiency evaluation: A non-parametric multilevel latent variable model approach. *Submitted*.

Gudeva, L. K., Dimova, V., Daskalovska, N., & Trajkova, F. (2012). Designing descriptors of learning outcomes for higher education qualification. *Procedia-Social and Behavioral Sciences, 46*, 1306–1311.

McMullen, J., & Hickendorff, M. (2018). Latent variable mixture models in research on learning and individual differences. *Learning and Individual Differences, 66*, 1–3.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.

Toomaneejinda, A. (2017). Zone of proximal development, dynamic assessment and learner empowerment. *LEARN Journal: Language Education and Acquisition Research Network, 10*(1), 176–185.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved threestep approaches. *Political Analysis, 18*(4), 450–469.

Considerations in Group Differences in Missing Values

Ambar Kleinbort, Anne Thissen-Roe, Rohan Chakraborty, and Janelle Szary

Abstract In recent years, employers are increasingly using artificial intelligence (AI) systems to summarize multidimensional psychological measurements to support employee selection decisions. It is important to evaluate and maximize the fairness of these AI models with regard to demographic groups such as gender and ethnicity. Different approaches have emerged, including obscuring group labels and maximizing the models' classification parity, neither of which guarantees a reduction in bias in operational settings (Corbett-Davies and Goel, The measure and mismeasure of fairness: A critical review of fair machine learning, 2018). The issue of fairness becomes more complex when missing measurements are imputed in the data used to train a model. The encoding of group differences can vary from the imputed data used for training, to complete real-world data. We tested how this can lead to unexpected observations of bias for the final model in production. To do this, we built debiased imputers that reduce group differences in missing values, and a paired, non-debiased version for each of them. We then built models on data imputed with each pair and tested their fairness with complete data sets labeled by groups (gender and ethnicity). We found that reducing the group differences encoded in imputed training data did not guarantee a more fair AI scoring model, and in some circumstances, it may result in a less fair model. We also found that alterations in missing data patterns post model building have little influence on fairness, and therefore note it's best to allocate more complete data to the training data set. Lastly, we were pleasantly surprised to see that neither of our imputation methods, when used to partially impute the demographic testing data sets, resulted in underestimations of groups differences.

Keywords Imputation · Missing data · Bias · Classification parity · Machine learning · Factor model imputation

A. Kleinbort (✉) · A. Thissen-Roe · R. Chakraborty · J. Szary
pymetrics, A Harver Company, Dallas, TX, USA
e-mail: ambar.kleinbort@harver.com; anne.thissen-roe@harver.com; rohan.chakraborty@harver.com; janelle.szary@harver.com

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_24

1 Background

At pymetrics, we do career matching using machine learning with data from a set of cognitive tasks. The contemporary career matching space puts a high emphasis on fairness, and the different ways it can be optimized and measured. In accordance with the US Equal Employment Opportunity Commission's guidelines, which govern recruiting laws in the US, the gold-standard measurement of practical significance of group differences in a selection method is minimum impact ratio. This is a ratio of the probability of success for the lowest and highest passing groups, which must be above 0.80. Furthermore, the guidelines say that the groups to be considered are gender and ethnicity (where each group exceeds 2% of the population from which employees are to be selected; United States Equal Employment Opportunity Commission, 1978).

With the entry of machine learning into the set of tools used for employee selection, concerns have been raised in both public and professional spaces: Will machine learning and AI tools replicate or even exacerbate patterns of historical discrimination in employment opportunity? Can these tools help to ameliorate historical discrimination, and if so, how? (Kassir et al., 2022; Corbett-Davies & Goel, 2018) We adopt an approach that emphasizes measurement, monitoring, and optimization of fairness outcomes at several points in our instruments' life cycle.

Once we have built an employee selection model, we use a holdout data set with demographics to measure expected impact ratios. However, once the model is deployed and we've collected real-world candidate data, the impact ratio often differs. In industry settings training data can differ from the real-world data acquired when a model is deployed in a number of ways. One common difference is that real-world data is often messier, with more missing values. Treatment of missing data has long been acknowledged to affect evaluations of fairness via self-selection effects (see, e.g.,Wainer, 1986); however, in our observations, the treatment of ignorable missing data (Rubin, 1987) matters as well.

Most frequently, a model is trained to fill in these missing values. Models with this goal are called imputers, and can be as simple as calculating the median or as complex as predicting values with a neural network. Imputing data can modify distributions and lead to overestimates or underestimates of true variability (Rubin, 1987). This can lead to faulty statistical and machine learning models. Further, we wanted to know how imputation of data can change the relationships between data distributions from demographic subgroups within the population, in linear or non-linear ways. These changes may appear as artifacts in any subsequent statistical or machine learning model built upon the imputed data, making it harder to establish or monitor selection model fairness.

We set out to test how distorting these feature distributions using imputation would impact the final fairness of our machine learning employee selection models. This was done by comparing the impact ratios obtained from a full testing data set to that of masked and imputed versions of that same data set. Furthermore, we developed debiased imputers, which avoid propagating or exacerbating group

differences in columns with missing data, and compared their effects on final model fairness to those of regular imputers.

2 Experimental Setup

In order to compare debiased and regular imputer performance, we built employee selection models for multiple jobs, and used a range of missing data percentages and missing mechanisms for data masking. We also tested two pairs of imputers, including Multiple Imputation by Chained Equations (MICE) and a factor model. As stochastic variation can impact multiple stages of the machine learning process, namely the masking of training data, model fitting, and masking of testing data, we performed replications of our experiment at each stage. We first tested the effect of changing the random seed used to mask the training data set that our machine learning models are fit on, using 25 seeds. Secondly, we iterated over 25 random seeds at the model fitting stage, meaning there was a single mask on the training data, but the model was fit 25 times. Lastly, we tried 25 different masks on the holdout set used to measure the impact ratio. For simplicity, we performed these replications for one missing type, occupation and imputer pair; we believe our results to be representative of what we would find in others.

2.1 Selection Models

In order to test how imputation affects model fairness in a comprehensive way, we created selection models for three distinct occupations: Human Resources, Finance, and Sales. For the purposes of this investigation, we used an adapted version of a soft-margin Support Vector Machine (SVM) to create the selection models. An SVM is a classification algorithm that seeks to find a hyperplane separating two classes. As described in Cortes and Vapnik (1995), it does so by minimizing

$$\frac{1}{2}\boldsymbol{W}^2 + C\sum_{i=1}^{n}\xi_i, \tag{1}$$

where W is the vector that is normal to the hyper-plane that separates the classes, C is a hyper-parameter and ξ is a slack variable.

As noted above, the minimum acceptable impact ratio for any of these models is 0.80, but for the purposes of this study we did not use any of our usual methods to improve model fairness, since we wanted to observe the effects of missingness and imputation alone. Thus, we observed impact ratios below 0.80 that would normally lead to model rejection and replacement later in our process.

2.2 Missingness Simulation

To simulate incomplete data for each model, we began with a full data set and removed data points using two different missingness mechanisms: Missing completely at random (MCAR) and missing at random (MAR). MCAR is characterized by data points being missing independently from both the observed and unobserved features, and occurs strictly at random. This means that MCAR patterns cannot be explained by the data that has been recorded nor by any other data. On the other hand, MAR data points can be accounted for in whole or part by the observed features, but are not influenced by unobserved values, including their own missing values (Rubin, 1976). We used an open source implementation by Muzellec et al. (2020) to produce MCAR and MAR masks for our data. Both were applied at 5, 10 and 20% missing (see appendix for code). These percentages were included because they are typical of what we have encountered in practice. Samples with greater than 20% missing are generally excluded from our models/analyses. We do observe MNAR mechanisms, but they are beyond the scope of this chapter.

2.3 Imputer Models

Once the missingness patterns were simulated, we built two pairs of imputers. First, by using MICE, which starts by filling in the missing values with the median. It then takes the feature that had the most missing values, and creates a regression to replace the filled in values with more accurate ones (based on the current values in all the other features, including the imputed values). At this point, it moves on to the feature with the second most missing values, does the same, and iterates through all the features until the values converge (Azur et al., 2011). To create the debiased version of the MICE imputer, we made each regression prioritize smaller errors for the minority group. Furthermore, the debiased MICE model was built separately for gender and ethnicity. We selected this model due to its optimal performance on our data (Chakraborty et al., 2022).

This means that for each regression

$$\hat{X}_n = \sum_{m \neq n}^{n} (\beta_{nm} * x_m) + error_n, \tag{2}$$

to obtain the vector of $\boldsymbol{\beta}_n$ coefficients, MICE uses the least squares difference

$$\hat{\boldsymbol{\beta}}_n = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \tag{3}$$

and the debiased MICE uses a weighted least squares difference where

$$\hat{\boldsymbol{\beta}}_n = \boldsymbol{w}^3 \left(\boldsymbol{X}^\top \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{4}$$

To get the demographic specific weights for our samples, $\boldsymbol{w}$, we start with G in categories^n, and let $\boldsymbol{w} = (n/c_1, n/c_2, ..., n/c_n)$ where $c_i = \Sigma_{g \in G} \delta_{G_i, g}$.

We also tested an imputer based on a factor model of our measures. This is, to our understanding, an unusual and perhaps novel imputation method. The factor model imputer begins by scoring each individual's factors by an exact method that ignores missing data (Thissen & Thissen-Roe, 2020). It proceeds to use point estimation, treating the factor scores as inputs to the regression equation form of the factor model, to fill in the missing values. That is, our multiple factor model for the vector of observed responses $\boldsymbol{y}_i$ for person i can be written as

$$\boldsymbol{y}_i = \boldsymbol{\Lambda} \boldsymbol{f}_i + \boldsymbol{\epsilon}_i, \tag{5}$$

in which the observations $\boldsymbol{y}_i$ and the vector of factor scores $\boldsymbol{f}_i$ are standardized, $\boldsymbol{\Lambda}$ is the matrix of factor loadings for $\boldsymbol{y}_i$ on $\boldsymbol{f}_i$ (used here as regression coefficients), and $\boldsymbol{\epsilon}$ is multivariate $N(\boldsymbol{0}, \boldsymbol{\Theta})$ in which $\boldsymbol{\Theta}$ is the variance-covariance matrix of the residuals. Then we can make a point estimate for the observations $\boldsymbol{y}_i$ as

$$\hat{\boldsymbol{y}}_i = \boldsymbol{\Lambda} \boldsymbol{f}_i, \tag{6}$$

and fill only the missing observations with their corresponding point estimates, leaving actual observations in place. (In our usual case, where the observations are not standardized, there is an additional transformation step on each side, in which the observed values are standardized prior to factor score estimation, and the estimated observations $\hat{y}_i$ are subjected to the reverse scale-and-shift operation to restore their original range of values. This could be built into the factor model; for practical reasons we do not.)

We had an existing theory of the factor structure of our measures, which we used as the base hypothesis for a pair of confirmatory factor analyses (CFA), one of which was constrained in ways designed to reduce group differences in the latent variables, accounting for gender and ethnicity simultaneously; that is, less group differences were encoded in the factor scores. This manipulation was verified to have achieved its goal; some factors had small group differences to begin with, but others had notable reductions. Between the two parallel CFA results, we had the materials for a debiased/regular matched pair of imputers, as we did with MICE.

In contrast to the debiasing method used for MICE, the strategy for reducing bias in the factor model was indirect. The target of bias reduction was group differences in the latent variable scores, not the subsequent imputed values. It was not clear a priori that the debiased factor model, when used as an imputer, would in fact result in less group differences in imputed data sets and subsequent models in the same way that the directly-debiased MICE imputer would. Therefore, we carried out the same empirical tests on the factor model imputer.

Table 1 Conditions for training and testing predictive models. This table is replicated for each occupation, imputer pair, at each missing percentage and mechanism

	Imputed test set	Full test set
Training set with debiased imputation	Debiased imputed set	Debiased full set
Training set with regular imputation	Regular imputed set	Regular full set

2.4 Impact Ratios

This set up resulted in four sets of impact ratio results per experiment, given that the imputed testing set and the full testing set are each passed through two models; one where the training set was imputed with the debiased imputation and one with the regular imputation. In each case, the same imputation model is used to produce the training set for model development and the test set for impact ratio evaluation. We did not measure how models behave when the training and testing sets are imputed differently, on the basis that this is unlikely in a production setting. The experimental design is summarized in Table 1.

3 Results

In varying the random seed across the steps of machine learning processes, we find that missing data and imputation affect subsequent fairness measures in distinct ways at each step. Firstly, we observed that changing the mask on the training set has a very heavy influence on final impact ratios, as can be seen by the large boxplots in Fig. 1a. At this stage, the missing data can flip which demographic groups pass at the lowest or highest rates. Changing the seed for the machine learning model also had an impact, as can be seen in Fig. 1b, but the order of the groups remained more stable. Moreover, we were very interested to see that varying missing data patterns beyond this point has very little effect on the impact ratio. This is depicted in Fig. 1c, where we see that changing the testing set mask creates very little variation. This suggests that in situations with limited availability of complete data, it is best to use as much of the complete data as possible in the training set, rather then trying to distribute it between the training and testing sets.

The resulting impact ratios for gender across criterion reveal that debiasing the imputation can lead to both better and worse parity amongst groups. This depends on the occupation, the missing percentages, and the missing mechanisms. As shown in Fig. 2, the models with debiased imputation (light shaded bars) alternate between having higher and lower impact ratios as compared to their counterpart models with regular imputation (dark shaded bars). This makes the overall difference of the mean impact ratio tend towards zero, at 0.004 (SE = 0.008).This suggests that minimizing the distortion of the relationships between distributions can sometimes be helpful

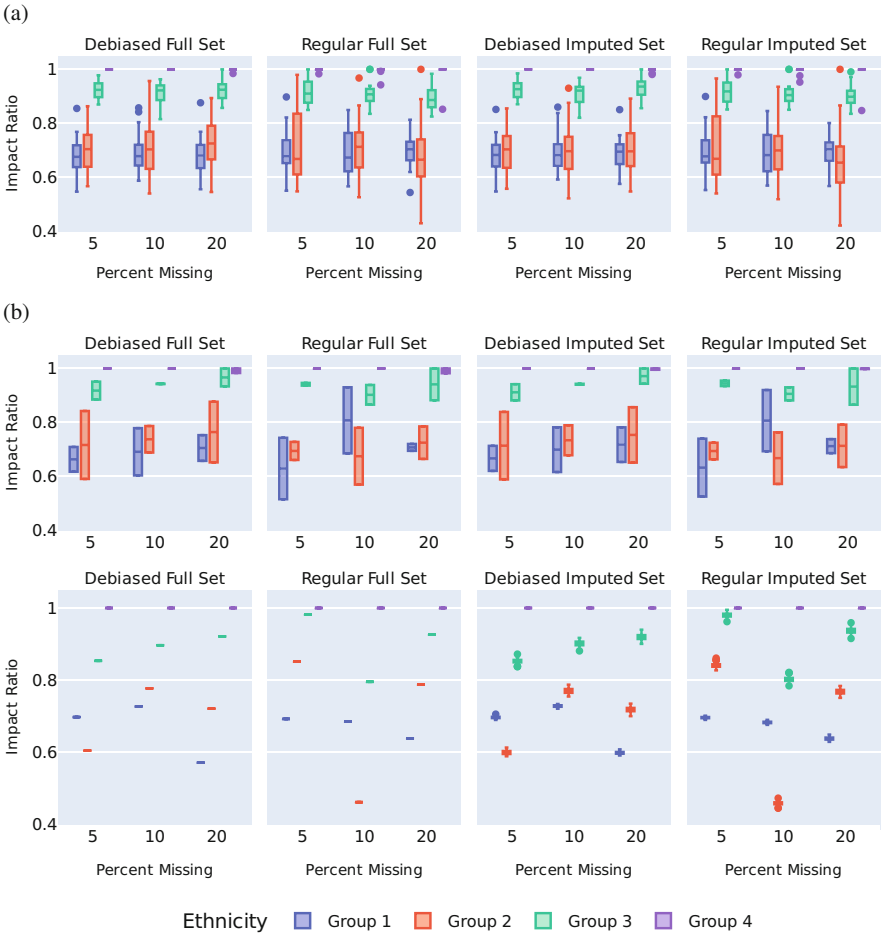


Fig. 1 Reliability of impact ratios for ethnicity groups. Each box represents the interquartile range of 25 repeated trials using different random seeds applied to the (**a**) training set masking, (**b**) model fitting, (**c**) testing set masking step for the Sales selection model, using MCAR simulation with MICE imputation

when optimizing model fairness, but it is not a guarantee due to the non-linear nature of the distortions.

Furthermore, we were pleasantly surprised to see that the expected (from the imputed testing set) and actual (from the full testing set) impact ratios had negligible differences, in spite of the changes in the imputed data. This can also be observed in Fig. 2, where we see that for any given set of conditions the impact ratios for models tested on imputed data (green bars) and full data (blue bars) are very similar. This holds true for the debiased and regular imputer conditions.

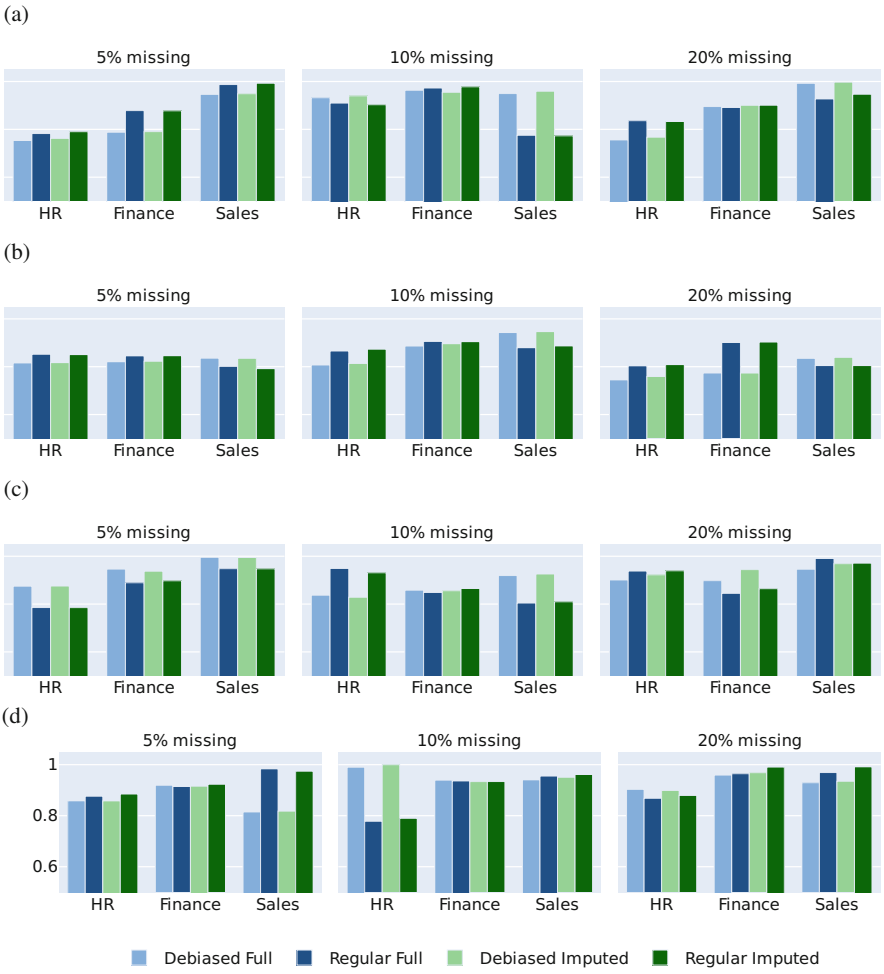


Fig. 2 Impact ratios for gender. Each row shows impact ratios obtained for a single trial of each imputer and missingness condition across various selection models and missing percentages. (**a**) MICE imputer, MCAR missingness. (**b**) MICE imputer, MAR missingness. (**c**) Factor model imputer, MCAR missingness. (**d**) Factor model imputer, MAR missingness

The impact ratios for ethnicity in Fig. 3 showed that debiased MICE imputers can either help or hurt overall fairness, which is demonstrated by similar impact ratios when aggregated across all other conditions (selection model, missing mechanism, and missing percentage). The difference of the means is 0.005 (SE = 0.012). Neither imputer affected the system enough to create large changes in the order of the ethnicity groups.

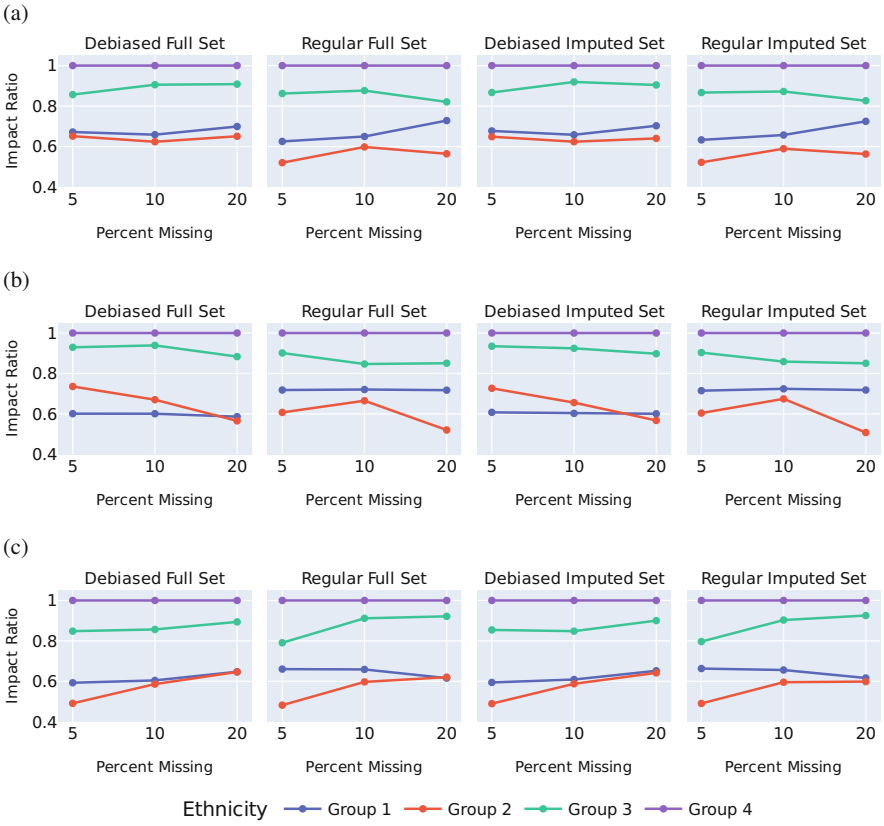


Fig. 3 Impact ratios for ethnicity with MICE imputation. Impact ratios obtained for a single trial of MICE imputation with MAR missingness across missing percentages for three different selection models. (**a**) Human resources model. (**b**) Sales model. (**c**) Finance model

Similarly, models trained with the debiased versus regular factor model imputers did not show a directional difference in overall fairness. For gender, the difference in the mean impact ratio is 0.005 (SE = 0.01), and for ethnicity the difference is 0.017 (SE = 0.017). However, debiasing these imputer models can have the effect of changing the order of demographic groups' pass rates across missing percentages (for a fixed occupation and missing mechanism). This effect is best demonstrated by the crossing lines in the factor imputer results in Fig. 4, contrary to the very mild flips in some of the MICE imputer results (see Fig. 3). As with the MICE results, we see that the expected versus actual impact ratios had minimal differences, as observed in how the imputed testing set results look similar to their full counterparts.

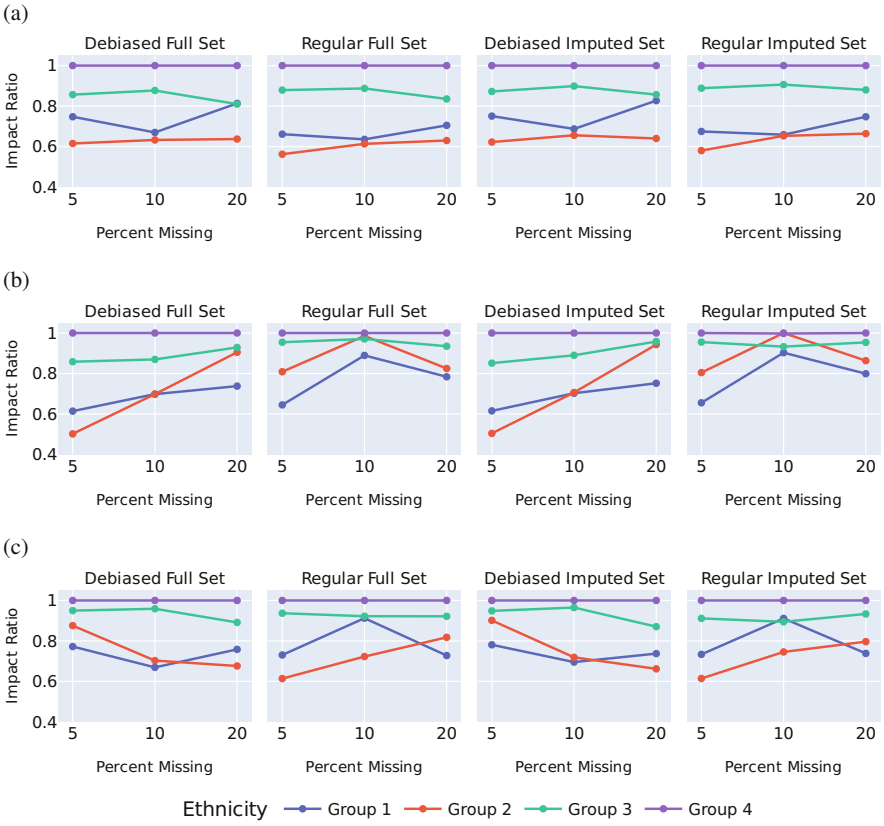


Fig. 4 Impact ratios for ethnicity with factor model imputation. Impact ratios obtained for a single trial of factor model imputation with MAR missingness across missing percentages for three different selection models. (**a**) Human resources model. (**b**) Sales model. (**c**) Finance model

4 Conclusion

In this study, we were able to establish that calculating an expected impact ratio from an imputed testing set gives a good approximation of actual impact ratios. We also showed that imputation technique has a heavy impact on final model fairness, which can be improved by debiasing the imputer. However, debiasing can also make models less fair. For this reason, the use of debiased imputers to optimize model fairness needs to be evaluated for each use case. In future studies, we will aim to explain the non-linear changes in the relationship between the distributions of each group to identify which scenarios will or will not benefit from debiased imputation. Additionally, we found that missing data has the heaviest impact on fairness at the training data set stage, and therefore recommend utilizing as much complete data as possible during this stage of model building. In further studies, we would also like to

see how debiasing imputers and concentrating the more complete data in the training set, as opposed to sharing it with the testing set, interacts with our usual methods for making models more fair. This will allow us to see how much improvement these methods can add to an already fairness-optimized workflow.

Acknowledgments We would like to thank Frida Polli, founder of pymetrics. We would also like to thank Amy Li and Kyle Mercury for their feedback throughout the project, and Mycchaka Kleinbort for his support. Furthermore, we thank everyone at IMPS 2022, especially the organizers, for bringing us together to share ideas.

Appendix: Code to Generate Missing Data Patterns

This code in Python represents how we simulated MCAR and MAR missingness patterns, as well as different missing percentages for each. The MNAR simulation details are excised since they were not utilized in this paper and are specific to the pymetrics data pipeline.

```
def produce_NA(p_miss, X, mecha = "MCAR",
    mecha_gen_type = "classic",opt = None, p_obs = None,
     trait_missing_probs_file = None,
    game_missing_probs_file = None,
    missing_trait_group_probs_file = None,
    game_trait_table = None,trait_year = '2020'):
     """
     Generate missing values for specifics missing-data
    mechanism and proportion of missing values.

     Parameters
     ----------
     X : torch.DoubleTensor or np.ndarray, shape (n, d)
    or pandas.DataFrame (n, d+...)
         Data for which missing values will be simulated
    .
         If a numpy array is provided, it will be
    converted to a pytorch tensor.
     p_miss : float
         Proportion of missing values to generate for
    variables which will have missing values.
     mecha : str,
             Indicates the missing-data mechanism to be
    used. "MCAR" by default, "MAR", "MNAR" or "MNARsmask
    "
     mecha_gen_type:
```

```
15        Indicates if the missing-data mechanism is:
16            - "classic"
17            - mimics our game_processors ("
    game_processor")
18     opt: str,
19        For mecha_gen_type == "classic" and mecha = "
    MNAR"
20        For mecha_gen_type == "game_processor" and
    mecha = "MAR", it indicates whether the missing-data
     mechanism is generated at a
21            - game level ("game_level")
22            - group trait level ("grouped_trait_level")
23     p_obs : float
24            If mecha_gen_type == "classic", and (mecha
    = "MAR" or ("MNAR" with opt = "logistic" or "quanti
    ")),
25            proportion of variables with *no* missing
    values that will be used for the logistic masking
    model.
26     trait_missing_probs_file: file path for pandas.
    DataFram,
27        File path for DataFrame containing trait level
    missing probabilities.
28        Must contain a column called 'missing_prob',
    indexed with missing trait corresponding each of
    those excess probabilities.
29     game_missing_probs_file: file path for pandas.
    DataFrame,
30        DataFrame containing game level missing
    probabilities.
31        Must contain a column called 'missing_prob',
    indexed with missing games corresponding each of
    those excess probabilities.
32     missing_trait_group_probs_file: file path for
    pandas.DataFram,
33        Dataframe containing excess probabilities of
    each of the groups of missing traits.
34        Must contain a column called 'excess_prob',
    indexed with missing traits corresponding each of
    those excess probabilities.
35     game_trait_table: dict,
36
37
```

```
        A map between games and their corresponding trait names.
    Returns
    ----------
    A dictionnary containing:
        'X_init': the initial data matrix.
        'X_incomp': the data with the generated missing values.
        'mask': a matrix indexing the generated missing values.
    """
    TRAITS = TRAITS2020 if trait_year == '2020' else TRAITS2022
    to_torch = torch.is_tensor(X) ## output a pytorch tensor, or a numpy array
    if not to_torch:
        X = X.astype(np.float32)
        X = torch.from_numpy(X)
    X = X.float()

    if mecha_gen_type == "classic":
        if mecha == "MCAR":
            mask = (torch.rand(X.shape) < p_miss).double()
        elif mecha == "MAR":
            mask = MAR_mask(X, p_miss, p_obs)
        elif mecha =='MNAR":
            raise ValueError("not implemented for this paper")

    X_nas = X.clone()
    X_nas[mask.bool()] = np.nan

    if not to_torch:
        return {'X_init': X.double().numpy() , 'X_incomp': X_nas.double().numpy(), 'mask': mask.numpy().astype('float')}
    else:
        return {'X_init': X.double() , 'X_incomp': X_nas.double(), 'mask': mask}
```

References

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research, 20*, 40.

Chakraborty, R., Kleinbort, A., Thissen-Roe, A., & Szary, J. (2022). How real is synthetic missing data? Impact of missing pattern modeling in imputer evaluation. *Joint Statistical Meetings*, Washington, DC, USA.

Corbett-Davies, S., & Goel, S. (2018). *The measure and mismeasure of fairness: A critical review of fair machine learning*. Preprint: arXiv:1808.00023.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*. https://doi.org/10.1007/bf00994018.

Kassir, S., Baker, L., Dolphin, J., & Polli, F. (2022). AI for hiring in context: a perspective on overcoming the unique challenges of employment research to mitigate disparate impact. *AI Ethics*. https://doi.org/10.1007/s43681-022-00208-x

Muzellec, B., Josse, J., Boyer, C., & Cuturi, M. (2020). *Missing data imputation using optimal transport*. arXiv. Retrieved from https://arxiv.org/abs/2002.03860

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581–592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

Thissen, D., & Thissen-Roe, A. (2020). Factor score estimation from the perspective of item response theory. In M. Wiberg, D. Molenaar, J. González, U. Bockenhold, & J.-S. Kim (Eds.), *Quantitative Psychology: 84th annual meeting of the Psychometric Society, Santiago, Chile, 2019* (pp. 171–184). Springer.

United States Equal Employment Opportunity Commission. (1978). *Uniform guidelines on employee selection procedures*.

Wainer, H. (1986). *Drawing inferences from self-selected samples*. Springer-Verlag.

Fully Latent Principal Stratification: Combining PS with Model-Based Measurement Models

Sooyong Lee, Sales Adam, Hyeon-Ah Kang, and Tiffany A. Whittaker

Abstract Despite prominent potentials of randomized controlled trials (RCTs) with computer-based interventions, log data poses a challenge since it differs in structure and size from the type of data commonly encountered in studies of causal mechanisms. The current study developed a method for RCTs suitable for big data, or complex implementation data. The method is an extension of principal stratification (PS), a causal framework used for studying how treatment effects vary as a function of post-treatment or intermediate variables. To exploit the complex structure of the log data, the proposed method can incorporate latent variables or measurement models into PS, substantially extending the scope of PS modeling into scenarios with multivariate and complex implementation data. With the method development, we did a simulation study to evaluate if our proposed FLPS model worked properly under various conditions, including sample sizes, number of items, effect sizes, and response rates, with different IRT models. FLPS models will allow researchers to gain deeper and more nuanced insights into the relationship between the effectiveness of the interventions under study, and how they are used. This ability will, in turn, deepen our understanding of our rapidly-evolving and growing interaction with technology, guiding the development of more effective interventions and guiding the implementation decisions of users.

Keywords Principal stratification · Randomized control trials · Latent variables · Causal inference · Log data

S. Lee (✉) · H.-A. Kang · T. A. Whittaker
The University of Texas at Austin, Department of Educational Psychology, Austin, TX, USA
e-mail: sooyongl09@utexas.edu

S. Adam
Worcester Polytechnic Institute, Worcester, MA, USA

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_25

1 Introduction

The data revolution in education has led to more data collection within randomized controlled trials (RCTs) to study program effectiveness. It is particularly the case in RCTs evaluating computer-based interventions, which allow researchers and administrators to collect implementation data in the form of log or clickstream data. For example, Pane et al. (2014) conducted a large-scale RCT to study the effectiveness of using computers in math learning. This study extensively gathered not only primary data for evaluating program effectiveness, but also log data obtained while using CTA1,[1] such as students' performance on each worked section, timestamps, and hints requested, to name a few. Such log data from technology RCTs presents an unprecedented opportunity for researchers to use the fine-grained and rich data to help understand how and why online interventions work.

Despite the prominent potential of RCT with computer-based interventions, log data poses a challenge since it differs in structure and size from the type of data commonly encountered in studies of causal mechanisms. Unlike the *implementation* or mediation data that most current statistical techniques are designed for, computer log data is highly multivariate, multilevel, and messy. Taking Pane et al.'s study as an example again, their data were gathered from 147 schools, spanning 52 diverse school districts in seven states over the course of two school years. In addition, log data was available—including, e.g., correctness and hint usage for each problem each student attempted—but only for students assigned to the treatment group who had access to the software. Furthermore, some log data, including the number of problems, widely varied between students in the treatment group (Sales & Pane, 2019). As a result, traditional statistical methods developed for causal models do not adapt easily to computer log data.

To exploit the complex log data structure, fully-latent Principal Stratification (FLPS) was proposed, incorporating a latent-variable-based measurement component into the broader causal model. The technique is an extension of principal stratification (PS), a causal framework used for studying how treatment effects vary as a function of post-treatment or intermediate variables (Feller et al., 2016; Frangakis & Rubin, 2002; Sales and Pane, 2019). Such FLPS modeling opens the door to a new frontier in causal modeling—combining PS and other causal frameworks with model-based measurement models, such as item response theory. The proposed method can incorporate continuous latent variables as measurement models into PS, substantially extending the scope of PS modeling into scenarios with multivariate and complex implementation data.

[1] The Cognitive Tutor Algebra1 (CTA1) is a software for intelligent tutoring systems that provides personalized learning for students through adaptive testing in math learning.

2 Classical PS

The classical PS is briefly reviewed to set the stage for FLPS. Principal stratification (Frangakis & Rubin, 2002) is based on the potential outcomes framework, in which the causal effect in an outcome Y between treatment assignments (Treatment or Control) is defined as the difference between two potential outcomes ($\tau_i \equiv Y_{T_i} - Y_{C_i}$). Randomized treatment assignment allows estimation of average effects for overall and baseline subgroups. However, variation in effects due to a variable M defined subsequent to treatment assignment may not have a causal interpretation.

In the PS framework, "principal" effects are estimated within strata (groups) of individuals with the same *potential* values of M, M_C and M_T. That is, principal effects are defined as $E[\tau|M_T = m_t, M_C = m_c] = E[Y_T - Y_C|M_T = m_t, M_C = m_c]$ (Frangakis & Rubin, 2002). When M itself is a feature of the intervention, such as in the RCT which the current study focuses on, M is only defined for the treatment group as M_T. In those cases, we can define "principal effects" for subgroups based on M_T (instead of M) as: $E[\tau|M_T = m] = E[Y_T|M_T = m] - E[Y_C|M_T = m]$. In other words, the principal effect is a special type of subgroup or moderation effect—the effect of treatment assignment among subjects who would, if assigned to treatment, implement the program in a particular way.

In the classical PS framework, M is a single measurement without error. When the intermediate variable M is multivariate—say, including measurements of students' master skills—classic PS models may use aggregated measures (e.g., the sample mean $\hat{m}$) to stratify on its potential values ($\hat{m_{T_i}}$). This approach, however, ignores measurement error in the aggregate as well as other relevant aspects of the measurement structure, and can produce misleading results.

3 Fully Latent PS

We propose "Fully Latent PS" or "FLPS," extending classical PS to model implementation data that includes several measurements, $m_i \equiv \{m_{i1}, .., m_{ij_i}\}$, where j_i denotes the number of measurements for subject i. FLPS incorporates the measurement process into the classical PS, specifying a distribution for the measurements, $p(m_i|\eta_{T_i})$, where η_T is a subject-level latent trait variable measuring the construct of interest. The T subscript of η_T denotes *potential* implementation; even though measurements m are only available for subjects in the treatment group, η_T is well-defined for all subjects in the experiment. It measures subjects's potential implementation—how they would implement the intervention if assigned to the treatment condition. The causal estimand in FLPS is $\tau(\eta_T) \equiv E[\tau|\eta_T]$—the averaged treatment effect for subjects who would, if assigned to treatment, implement the interventions as η_T. Unlike a classical PS intermediate variable M_T, latent variables such as η_T are not observed in either the treatment or the control group of the study. Thus, distributions for both treated and control potential

outcomes follow mixture distributions. For instance, when η_T is continuous, we may model outcomes as

$$p(Y_T|m,x) = \int p(Y_T|\eta_T,x)p(m|\eta_T)p(\eta_T|x)d\eta_T \text{ and } p(Y_C|x)$$
$$= \int p(Y_C|\eta_T,x)p(\eta_T|x)d\eta_T, \quad (1)$$

where x is a vector of covariates and $p(\eta_T|x)$ is a model for η_T as a function of x. Although η_T is unobserved and must be estimated in both the treatment and control groups, the data differ markedly between the two groups: the model for η_T in the treatment group includes both measurements m and covariates, whereas in the control group, only covariates are available. We took a Bayesian approach to estimation, with the goal of estimating the posterior distribution of model parameters θ (Gelman et al., 1995), which can be computed as:

$$p(\theta|Y,Z,X,m_{i:Z_i=1})$$
$$\propto p(\theta) \times \prod_{i:Z_i=1} \int p(Y_i|Z_i,\eta_{T_i},X_i;\theta)p(m_i|\eta_{T_i};\theta)p(\eta_{T_i}|X_i,\eta)d\eta_{T_i} \times \quad (2)$$
$$\prod_{i:Z_i=0} \int p(Y_i|Z_i,\eta_{T_i},X_i;\theta)p(\eta_{T_i}|X_i;\theta)d\eta_{T_i},$$

where $p(Y_i|Z_i,\eta_{T_i},X_i;\theta)$ is a model as a function of covariates X and latent parameters η_Y for treatment or control potential outcomes when $Z_i = 0$ or 1, respectively. $p(m_i|\eta_T;\theta)$ is a measurement model. $p(\eta_{T_i}|X_i;\theta$ is a model for η_T as a function of X. $p(\theta)$ is the prior probability density function regarding the FLPS parameters.

Sales and Pane (2019) explored Rasch measurement modeling in FLPS to model mastery data. In this study, we extend the framework to accommodate other item response models, such as the two-parameter logistic (2PL) model (Birnbaum, 1968), the generalized partial credit model (GPCM; Muraki, 1997; Masters, 2016) and the graded response model (GRM; Samejima, 1969). Below we present performance of the extended FLPS in simulation settings.

4 Simulation Study

A Monte Carlo simulation study was conducted to investigate FLPS models' operating characteristics. The simulation was designed to mimic RCT implementation data gathered during the RCT of Pane et al. (2014). All simulation studies were carried out using R version 3.5.1 (R Core Team, 2021) and Stan (Stan Development Team, 2016).

4.1 Design

The manipulated simulation factors include sample size (N), the number of items (J), and four item response models (Rasch, 2PL, GPCM, and GRM). Fixed factors are as follows: Each student in the treatment group attempted a random 60% of available items, whereas the students in the control group did not attempt any items.[2] Two covariates were included in the model, which explained 50% of the variance in the latent factor. 20% of the variance of the outcome is accounted for by the treatment assignment, covariates, and the latent factor. Table 1 summarizes details on the parameterization of these factors. The strength of the relationship between Z and Y (τ_0); between η_T and $Y_T - Y_C$ (τ_1); between the latent variables (η_T) and potential outcomes (Y) (ω); between X and η_T (β); between Y and η_T (γ) was randomly drawn from the uniform distribution with the range differing by relationships.

In terms of measurement models, intercept parameters were from the standard normal distribution for the Rasch and 2PL models. The items for the GPCM and GRM have four categories with the three intercepts drawn from the uniform distribution with a minimum distance of 0.5 to ensure enough space between the intercepts. The mean-centered values were used for intercepts of polytomous

Table 1 Description of the simulation study

Condition	Simulation factors	Values	Notation
Manipulated	Measurement model	Rasch, 2PL, GPCM, GRM	Model
	Sample size	500, 1000, 2000	N
	Number of items	50, 100, 200	J
Fixed	Number of covariates	2	
	Percentage of items administered	0.6	
	Strength of relationship		
	η_T and Y_C	$U(0.1,0.3)$	ω
	η_T and $Y_T - Y_C$	$U(-0.2, -0.1)$	τ_1
	Z and Y	$U(0.2,0.4)$	τ_0
	Predictive power of η	0.5	
	Predictive power of Y	0.2	
	Measurement model parameters		
	Intercept	$N(0,1)$ for binary data	d
		$U(0.5,1)$ for polytomous data	
	Slope	$LogN(0.1,1.3)$	a

Note. The number of items (J) was fixed at 100 when evaluating the performance under different calibration sample sizes. Similarly, the sample size (N) was fixed at 1000 when evaluating the performance under different item set sizes

[2] Each treatment subject attempted 30, 60, and 120 items out of the total number of items (50, 100, and 200).

models. For all of the models, item slope parameters were generated from the log-normal distribution with a mean of 0.1 and a standard deviation of 0.3.

The FLPS model was analyzed in Stan via the 'rstan' package in R. For priors, we applied the log-normal prior distribution for the slope parameters and the normal distribution for the intercepts for the measurement models.[3] For the structural model, the priors for all the parameters were uniformly distributed. For the Bayesian estimation, two MCMC chains with 5000 iterations each were used to estimate the posterior distributions, with the first 2000 samples discarded in the burn-in period. The mean of the posterior distribution for each parameter was taken as the MCMC estimates.

4.2 Evaluation

For evaluating the estimation accuracy, we examined bias and root mean squared error (RMSE) of the final estimates. In addition to the distance measures, we also examined coverage rate of the credible interval estimates to evaluate fidelity of standard error estimates. The credible interval was obtained as the 2.5th and 97.5th percentiles of the posterior probability distribution of each estimand. The coverage rate evaluates the average proportion of times the credible interval includes the generating parameter.

5 Results

5.1 Recovery of Measurement Model Parameters

Table 2 presents bias, RMSE, and coverage rates of the measurement model parameter estimates observed under the different sample-size and measurement-size conditions. The measurement model parameter estimates on the whole showed adequate accuracy and precision. The error statistics were reasonably small (average bias = 0.072, RMSE = 0.335) and the precision of the uncertainty estimate was maintained adequately high (average coverage rate = 0.916).

Under the different sample-size conditions, the trends related to the sample size were generally consistent with the expectation—the larger the samples, the more accurate the estimates were. With the bias stable around zero, the values of RMSE declined from 0.300 to 0.229 and 0.189 with increases in the sample size. The results revealed that for sample sizes of 500 or fewer, the coverage rates were around the

[3] In terms of intercept priors, the standard normal prior was applied to the 2PL, whereas normal priors with −1, 0, and 1 as the mean and 1 SD were applied to each of the three intercepts for the polytomous models.

Table 2 Recovery of measurement model parameters across conditions

			Bias			RMSE			Coverage		
Model	N_T	J	a	d	η_T	a	d	η_T	a	d	η_T
Rasch	250	100	–	0.008	−0.004	–	0.213	0.334	–	0.948	0.948
	500	100	–	0.005	−0.004	–	0.152	0.330	–	0.956	0.949
	1000	100	–	0.004	−0.003	–	0.109	0.327	–	0.955	0.951
	500	50	–	0.006	−0.005	–	0.154	0.436	–	0.950	0.951
	500	100	–	0.005	−0.004	–	0.152	0.330	–	0.956	0.949
	500	200	–	−0.003	0.002	–	0.151	0.243	–	0.948	0.950
2PL	250	100	0.021	0.004	0.018	0.321	0.239	0.420	0.936	0.952	0.952
	500	100	0.001	0.005	0.010	0.220	0.178	0.369	0.937	0.946	0.951
	1000	100	0.028	0.004	−0.003	0.153	0.125	0.333	0.957	0.952	0.951
	500	50	0.032	0.001	0.007	0.241	0.176	0.458	0.929	0.956	0.949
	500	100	0.001	0.005	0.010	0.220	0.178	0.369	0.937	0.946	0.951
	500	200	0.009	−0.005	0.018	0.222	0.169	0.299	0.936	0.953	0.948
GPCM	250	100	0.043	−0.002	0.009	0.252	0.422	0.300	0.933	0.874	0.947
	500	100	0.007	0.003	0.015	0.164	0.209	0.259	0.945	0.957	0.955
	1000	100	0.003	0.001	0.006	0.125	0.151	0.238	0.929	0.956	0.949
	500	50	−0.004	−0.004	0.007	0.166	0.212	0.320	0.934	0.956	0.955
	500	100	−0.007	0.003	0.015	0.164	0.209	0.259	0.945	0.957	0.955
	500	200	−0.015	0.000	0.013	0.160	0.241	0.214	0.944	0.936	0.950
GRM	250	100	0.021	−0.005	0.014	0.252	0.201	0.345	0.930	0.958	0.951
	500	100	0.005	−0.001	0.011	0.180	0.145	0.309	0.927	0.959	0.948
	1000	100	−0.004	−0.001	0.007	0.127	0.104	0.287	0.924	0.959	0.950
	500	50	0.000	−0.003	0.009	0.183	0.147	0.393	0.921	0.959	0.953
	500	100	0.005	−0.001	0.011	0.180	0.145	0.309	0.927	0.959	0.948
	500	200	−0.004	−0.005	0.007	0.166	0.144	0.244	0.944	0.957	0.953

Note. N_T: Size of the treatment group. a: Slope parameter of the item response model. d: Intercept parameter of the item response model. η_T: Latent trait score of the treatment group. The trait estimates of the control group subjects showed average bias of 0.002, RMSE of 1.021, and coverage rate of 0.922. The number of measurement items was fixed at 100 throughout

nominal value of 0.95, regardless of the type of measurement models, except for the GPCM intercept estimates with $N = 250$. The low coverage in the GPCM is partly due to the prior for the intercepts under such a small sample size. The location of the intercept priors for the GPCM shifted the estimates for the first and third intercepts, which led to low coverage. As expected, longer assessments entailed more precise trait recovery under the different sample size conditions. As the number of items increased from 30 to 60 and 120, the bias of the trait estimates stayed around zero (0.03 on average) while RMSE decreased from 0.269 to 0.229 and 0.205 on average. The coverage rate of the interval estimates was kept at the nominal level, averaging 0.951 rate.

5.2 Parameter Recovery for Structural Models

Table 3 presents the results of the structural paths (the difference in Y between Z (τ_0), the principal effect (τ_1), the effect of η on Y (ω), the effect of X on η (β), and the effect of X on Y (γ) in terms of the bias, RMSE, and coverage rates. Overall, the results of the structural model estimates showed that bias and the RMSE were small, and coverage rates for structural paths were well above 0.9. Under different sample sizes, overall, the RMSE values associated with all the structural model parameters were close to zero, falling below 0.1, regardless of the simulation condition. The larger the sample size, the smaller the RMSE became for each parameter. Across all of the measurement models, the RMSE ranged from 0.04 0.09 with $N = 500$ and ranged from 0.02 to 0.05 with $N = 2000$. In contrast to the sample-size conditions, the values of RMSE associated with the structural parameters showed no noticeable difference between the varying number of items. The differences were less than 0.01 in RMSE. However, with $N = 1000$, the coverage rates are all above 0.9 across the number of items.

Table 3 Recovery of structural model parameters across conditions

			RMSE					Coverage				
Model	N	J	τ_0	τ_1	ω	β	γ	τ_0	τ_1	ω	β	γ
Rasch	500	100	0.066	0.067	0.080	0.069	0.041	0.970	0.940	0.950	0.945	0.930
	1000	100	0.046	0.048	0.055	0.054	0.028	0.920	0.950	0.960	0.930	0.965
	2000	100	0.031	0.030	0.037	0.036	0.020	0.960	0.970	0.970	0.960	0.945
	1000	50	0.045	0.044	0.064	0.051	0.026	0.960	0.970	0.920	0.960	0.925
	1000	100	0.046	0.048	0.055	0.054	0.028	0.920	0.950	0.960	0.930	0.965
	1000	200	0.042	0.044	0.057	0.049	0.030	0.970	0.950	0.930	0.935	0.935
2PL	500	100	0.070	0.079	0.090	0.094	0.041	0.970	0.940	0.940	0.945	0.970
	1000	100	0.046	0.051	0.060	0.065	0.028	0.970	0.970	0.960	0.955	0.925
	2000	100	0.035	0.040	0.048	0.040	0.021	0.950	0.940	0.940	0.975	0.970
	1000	50	0.047	0.059	0.071	0.057	0.028	0.970	0.930	0.940	0.970	0.960
	1000	100	0.046	0.051	0.060	0.065	0.028	0.970	0.970	0.960	0.955	0.925
	1000	200	0.043	0.055	0.065	0.054	0.027	0.980	0.910	0.950	0.935	0.945
GPCM	500	100	0.059	0.072	0.073	0.077	0.040	0.980	0.980	0.990	0.955	0.955
	1000	100	0.052	0.044	0.051	0.048	0.028	0.910	0.940	0.990	0.960	0.955
	2000	100	0.033	0.036	0.047	0.033	0.019	0.940	0.940	0.890	0.955	0.955
	1000	50	0.051	0.056	0.065	0.055	0.028	0.930	0.930	0.910	0.970	0.935
	1000	100	0.052	0.044	0.051	0.048	0.028	0.910	0.940	0.990	0.960	0.955
	1000	200	0.049	0.052	0.064	0.048	0.030	0.950	0.920	0.910	0.960	0.945
GRM	500	100	0.064	0.080	0.093	0.074	0.041	0.940	0.910	0.930	0.965	0.930
	1000	100	0.048	0.046	0.057	0.057	0.025	0.970	0.940	0.970	0.955	0.960
	2000	100	0.034	0.039	0.044	0.040	0.018	0.940	0.930	0.950	0.935	0.950
	1000	50	0.047	0.049	0.054	0.058	0.026	0.940	0.950	0.950	0.955	0.975
	1000	100	0.048	0.046	0.057	0.057	0.025	0.970	0.940	0.970	0.955	0.960
	1000	200	0.051	0.051	0.059	0.048	0.028	0.940	0.970	0.960	0.960	0.955

Note. N: Sample size; J: The number of items

6 Summary and Future Direction

The purpose of this study was to propose a flexible PS framework that incorporates various measurement models. The framework can accommodate large, complex implementation data that have been commonly observed in RCTs. Our simulation study showed that the proposed FLPS works properly with different item response models. The results demonstrated some promising potential for causal inference. The FLPS framework allows researchers to gain deeper and more nuanced insights into the relationship between the effectiveness of the interventions under study—in particular when big implementation data is available, such as in evaluations of computer-based interventions. Moreover, the FLPS framework will facilitate the thorough study of causal mechanisms in various measurement designs, permitting subsequent study designs to be optimized to reduce measurement error.

We conclude the paper with the limitations of the current study and possible future directions. The current study considered the item response models as a starting point. Future study can consider more flexible latent variable models such as factor analytic model, latent class models, or factor mixture models. Future study can also consider more complex relations in the log-data, such as multidimensional latent structures. Furthermore, this study limited the well-specified measurement models. In this simulation study, the population models were used as the measurement models for FLPS. In practice, however, it is not always the case. One might use misspecified measurement models mistakenly or for convenience. Future research can investigate how the FLPS could be robust with a misspecified measurement model.

Acknowledgments The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D210036. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Appendix

```
# Install the package for data generation
# devtools::install_github("sooyongl/IMPS2022")
library(IMPS2022)
library(rstan)

# Generate data
sdat <- makeDat(
  N = 200,        # number of total sample (Half is
  for the treatment group)
  R2Y = 0.2,      # Proportion of outcome explained by
   covariates
  R2eta = 0.5,    # Proportion of latent variance
```

```
  explained by covariates
  omega = 0.5,    # Effect of latent variable on
  outcome
  tau0 = 0.3,     # Difference in outcome between
  Treatment and Control
  tau1 = -0.15,   # Principal effects
  lambda = 0.5,   # Missingness on item responses
  nsec = 20,      # Number of items
  lvmodel = '2pl' # Latent variable models
  )

# Write the Stan code
# Other stan scripts can be found in sooyongl/IMPS2022,
 or loaded by mk_stan(\grm") for example.
stan_code <- "
data{
//Sample sizes
  int<lower=1> nsecWorked;
  int<lower=1> ncov;
  int<lower=1> nstud;
  int<lower=1> nsec;
  int<lower=1> nfac;
  int<lower=0> min_k;
  int<lower=1> max_k;

// Data indices
  int<lower=1,upper=nstud> studentM[nsecWorked];
  int<lower=1,upper=nsec> section[nsecWorked];

// Index for factor loadings
  matrix[nsec, nfac] factoridx;
  int<lower=0> firstitem[nsec];

// Input data
  int<lower=min_k,upper=max_k> grad[nsecWorked];
  matrix[nstud,ncov] X;
  int<lower=0,upper=1> Z[nstud];
  real Y[nstud];
}

parameters{
  // IRT model
  vector[nstud] eta;  // Latent traits
  real<lower=0> sigU; // Latent variable variance

  matrix<lower=0, upper=10>[nsec, nfac] a1_free; //
  Item Slopes
  real d[nsec];                 // Item intercepts

  vector[ncov] betaU; // Covariate effects on latent
  variable
  vector[ncov] betaY; // Covariate effects on outcome

  real omega;
```

```
  real yint;
  real tau0;
  real tau1;

  real<lower=0> sigY[2]; // Outcome variance
}

transformed parameters {
  matrix<lower=0, upper=10>[nsec, nfac] a1;

  // Factor loading constraints
  for(jjj in 1:nfac) {
    for(jj in 1:nsec) {
      if(factoridx[jj, jjj] != 0) {
        if(firstitem[jj] == 1) { // first loading per
        factor constrained to 1.
        a1[jj, jjj] = 1;
        } else {
          a1[jj, jjj] = a1_free[jj, jjj];
        }
      } else {
        a1[jj, jjj] = 0;
      }
    }
  };
}

model{
  vector[nstud] muEta;
  vector[nstud] muY;
  real sigYI[nstud];

// Fully Latent Principal Stratification model
// Structural part -----------------
   for(i in 1:nstud){
     muEta[i] = X[i, ]*betaU;
 muY[i]  = yint+ omega*eta[i] + Z[i] * (tau0 +
tau1*eta[i]) + X[i,]*betaY;
 sigYI[i]=sigY[Z[i]+1];

 eta[i] ~ normal(muEta[i], sigU);
 Y[i] ~ normal(muY[i], sigYI[i]);
   };

// Measurement part -----------------
   for(j in 1:nsecWorked) {
     grad[j] ~ bernoulli_logit(d[section[j]] +
     a1[section[j],1] * eta[studentM[j]]);
   };

// Priors ------------------
// IRT priors
   d ~ normal(0, 1);
   for(i in 1:nsec) {
```

```
      for(j in 1:nfac) {
        a1_free[i, j] ~ lognormal(0, 1);
      };
    };

// Priors for structural model
  betaY ~ normal(0, 1);
  betaU ~ normal(0, 1);
  omega ~ normal(0, 1);
  yint ~ normal(0, 1);
  tau0  ~ normal(0, 1);
  tau1 ~ normal(0, 1);
}
"

# Run a FLPS model
fit <- rstan::stan(model_code = stan_code, data =
 sdat$stan_dt)
summary(fit)
```

References

Birnbaum, A. (1968) Some Latent trait models and their use in inferring an examinee's ability. In: Lord, F.M. and Novick, M.R., eds., *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, 397–479.

Feller, A., Greif, E., Ho, N., Miratrix, L., & Pillai, N. (2016). *Weak separation in mixture models and implications for principal stratification*. Preprint. arXiv:1602.06595.

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

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Masters, G. N. (2016). Partial credit model. In W. J. van der Linden (Ed.), *Handbook of item response theory, Volume one* (pp. 137–154). Chapman and Hall/CRC.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–164). New York, NY: Springer New York.

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of cognitive tutor algebra i at scale. *Educational Evaluation and Policy Analysis, 36*(2), 127–144.

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Sales, A. C., & Pane, J. F. (2019). The role of mastery learning in an intelligent tutoring system: Principal stratification on a latent variable. *The Annals of Applied Statistics, 13*(1), 420–443.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4), 100.

Stan Development Team. (2016). *RStan: The R interface to Stan*. R package version 2.17.3.

Multilevel Reliabilities with Missing Data

Minju Hong and Zhenqiu Laura Lu

Abstract Reliabilities are widely used in social and behavioral sciences. The main purpose of this study was to investigate the performance of reliabilities for multilevel data with missing values. We examined the accuracy and convergence of multilevel reliabilities with missing values. The single-level reliabilities were also compared. In the simulation study, we considered different conditions, including missing data mechanisms, missing data techniques, missing data proportions, sample sizes at multilevel levels, and intra-class correlations. Results showed that, in general, multilevel reliabilities performed better than single-level reliabilities. Regarding missing data techniques, list-wise deletion method is not recommended.

Keywords Reliability · Missing data · Multilevel confirmatory factor analysis

1 Introduction

Reliability is one of the important features of a measurement tool such as a test or a questionnaire. Under the factor analysis framework (Jöreskog, 1971), we can estimate reliabilities by treating the true scores and the error as the latent variables. There are three types of measurement models to explain the relationship between factors and indicators: parallel, tau-equivalent, and congeneric. Because the assumptions for the parallel model are too strict in practical testing situations (Crocker & Algina, 1986), reliabilities have been estimated based on either the tau-equivalent or the con-generic model. The most common reliability estimate is the coefficient alpha (Cronbach, 1951). However, it is controversial to use the coefficient alpha because the tau-equivalent assumption could be violated in actual

M. Hong (✉)
University of Arkansas, Fayetteville, AR, USA
e-mail: minjuh@uark.edu

Z. L. Lu
University of Georgia, Athens, GA, USA

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_26

test administration situations (Cortina, 1993; Sijtsma, 2009). In addition, as test administration has become more complicated, traditional reliability estimates are more likely to lead to biased results in some cases such as when data have a multilevel structure (Bonito et al., 2012). Another problematic case involves the existence of missing values in the data set (Enders, 2003, 2004; Raykov, 2009).

Therefore, this study aims to investigate the performance of both the traditional (i.e., single-level) and multilevel (i.e., within-group level and between-group level) reliability estimates under a multilevel data structure with missing values. Based on literature, we consider the conditions of multilevel data structure (i.e., number of clusters, cluster sizes, and intra-class correlations) and the conditions of missing values (i.e., missing data mechanisms, missing data proportions, and missing data techniques).

2 Theoretical Backgrounds

2.1 Reliabilities for Single-Level Data

Based on the classical test theory, an observed score X has two components, the true score T and the error E:

$$X = T + E \tag{1}$$

Then, reliability ρ_{XT} is defined as the ratio of the true score variances σ_T^2 over the observed score variances σ_X^2 (Crocker & Algina, 1986):

$$\rho_{XT} = \frac{\sigma_T^2}{\sigma_X^2} \tag{2}$$

The well-known reliability Cronbach's alpha (coefficient alpha) is based on parallel or tau equivalent measurement. It is calculated by

$$\alpha = \frac{I^2 \overline{\sigma}_{ij}}{\sum \sigma_X^2} \tag{3}$$

where I is the total number of items in the test X, $\overline{\sigma}_{ij}$ ($i \neq j$ with $i, j = 1, \ldots, I$) is the average of off-diagonal elements, and $\sum \sigma_X^2$ is a sum of all elements of the observed scores' variance–covariance matrix. Another well-known reliability coefficient omega (McDonald, 1978) is for the congeneric measurement and is calculated by

$$\omega = \frac{\left(\sum_{i=1}^{I} \lambda_i\right)^2}{\left(\sum_{i=1}^{I} \lambda_i\right)^2 + \sum_{i=1}^{I} \theta_{ii}} \tag{4}$$

where λ_i is the factor loading and θ_{ii} is the error variance of the ith item in the test.

2.2 Reliabilities for Multilevel Data

The data in the field of education are often multilevel (Bryk & Raudenbush, 1987). Under the factor analysis model framework, Muthén (1994) applied the multilevel confirmatory factor analysis (MCFA) model to analyze the educational data set. Then, we can specify the two-level MCFA model to decompose the observed score into

$$X_{ikg} = X_{Big} + X_{wikg} = \left(\lambda_{Bi}\eta_{Bg} + \varepsilon_{Big}\right) + \left(\lambda_{wi}\eta_{wkg} + \varepsilon_{wikg}\right) \tag{5}$$

where X_{Big} is the component of the observed score at the between-group level, and X_{wikg} is the component at the within-group level.

Geldhof et al. (2014) proposed the level-specific reliability estimates; their idea was to apply coefficients alpha and omega to be estimated at each level of the data separately. If the test is tau equivalent, at the within-group and the between-group levels, we assume that factor loadings are the same, and error variances are varied among the ith test item. Then, level-specific alpha, including the within-group level coefficient alpha α_w and between-group level coefficient alpha α_B, are calculated by

$$\alpha_w = \frac{I^2\overline{\sigma}_{wij}}{\sum \sigma^2_{X_w}} \tag{6}$$

$$\alpha_B = \frac{I^2\overline{\sigma}_{Bij}}{\sum \sigma^2_{X_B}} \tag{7}$$

If the test is congeneric, that is, when we relax the assumptions of the same factor loadings, then, level-specific omega, that is within-group level coefficient omega ω_w and between-group level coefficient omega ω_B, are calculated by

$$\omega_w = \frac{\left(\sum_{i=1}^{I} \lambda_{wi}\right)^2}{\left(\sum_{i=1}^{I} \lambda_{wi}\right)^2 + \sum_{i=1}^{I} \theta_{wii}} \tag{8}$$

$$\omega_B = \frac{\left(\sum_{i=1}^{I} \lambda_{Bi}\right)^2}{\left(\sum_{i=1}^{I} \lambda_{Bi}\right)^2 + \sum_{i=1}^{I} \theta_{Bii}} \tag{9}$$

2.3 Reliabilities with Missing Values

Missing data are the unobserved values in the data set for many reasons (Little & Rubin, 2002). The previous studies showed that missing values in the data set are one of the most important factors for the reliability estimation (Enders, 2003, 2004; Raykov, 2009). Little and Rubin (2002) explained three missing data mechanisms. If the probability of missing data is not dependent on any factors of the entire data set, it is missing completely at random (MCAR). If the probability of missingness is dependent on the probability of the observed values in the data set, it is missing at random (MAR). If the probability of missingness is dependent on the probability distribution of the missing values in the data set, it is missing not at random (MNAR).

To handle the missing values, the most commonly used method is the listwise deletion (LD) method, which is the way to get rid of the cases having the missingness. However, under the MNAR, data deletion methods could cause biased results of statistical analysis, reduce the statistical power, and invalidate the conclusions of the study.

To overcome the limitation of LD, many researchers use either data imputation methods, which is a method to replace the missingness with the most plausible values, or the full information maximum likelihood (FIML) estimation method, which is a way to estimate the parameters from the incomplete data by augmenting the information from observed data and an underlying probability model. Based on literature, the conditions including the missing data mechanisms (MCAR, MAR, MNAR), the missing data proportions (up to 40%), the missing data techniques (LD method, FIML method, and data imputation methods such as person mean imputation, or item mean imputation), the population values of the reliability estimates, the sample sizes (100–3000 samples), and the test length (3–20 items) were the significant factors for the reliability estimation with missing values in the data set. (Deng & Chan, 2017; Edwards et al., 2021).

3 Methods

3.1 Data Generation

We first generated the complete data based on an MCFA model with one factor and six indicators, as shown in Fig. 1. In the within-group level model, the factor was

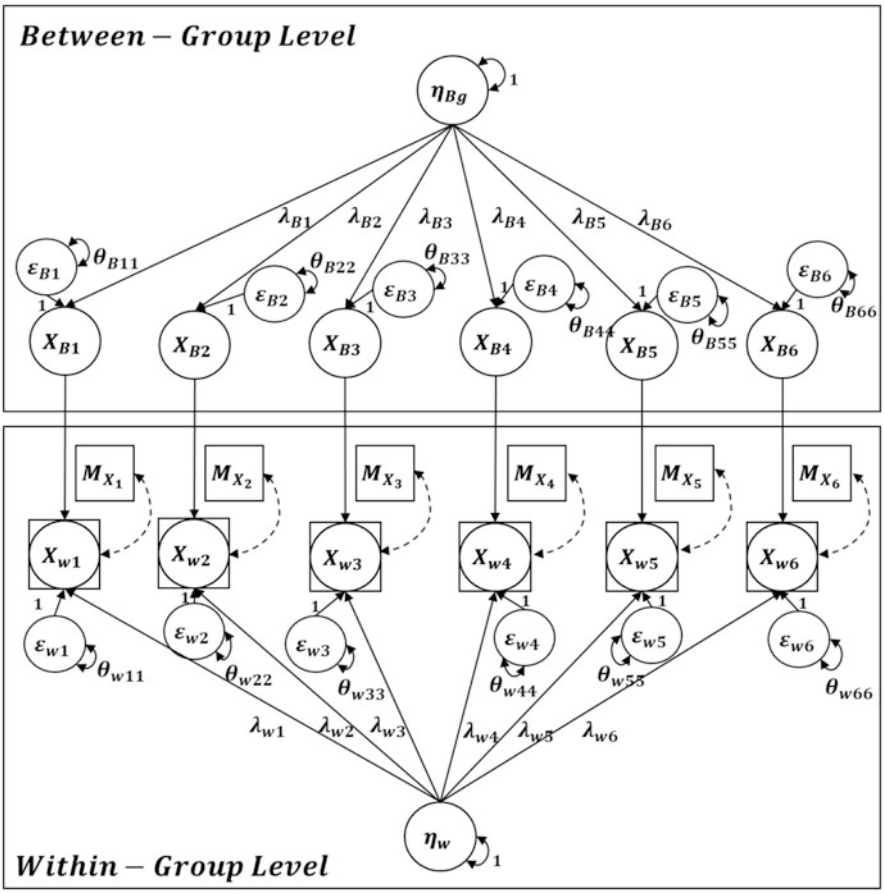


Fig. 1 Path diagram of the model for data generation and data analysis

shown as the circle in the lower plate with the factor score $\eta_w \sim N(0, 1)$. Factor loadings were λ_{wi} and the corresponding errors of each indicator were $\varepsilon_{wi} \sim N(0, \theta_{wii})$. And M_{X_i} in the square indicated whether the corresponding indicator X_i was observed with the dashed line representing the probability density of missing data $p\left(M_{X_i}\right)$.

At the between-group level, a single factor was represented as the circle of the factor score $\eta_B \sim N(0, 1)$ related to the six indicators with factor loadings λ_{Bi}. Different from the within-group level model, the between-group level indicators X_{Bi} were shown as the latent variables because X_{Bi} indicated the shared but unobservable features of the within-group level units belonging to the same between-group level unit. Their corresponding errors were $\varepsilon_{Bi} \sim N(0, \theta_{Bii})$ seen as the circles in Fig. 1.

To generate missing values, we used the procedure as follow: For MCAR, we created three binary variables corresponding to the indicators X_4–X_6 based on the

Bernoulli distribution with the probability p (i.e., the missing data proportion). If the value of the binary variable was 1, then the value of the corresponding indicator was deleted; if the value of the binary variable was 0, then we treated the value of the corresponding indicator as the observed value. For MAR, we matched three pairs of the six indicators selected as (X_1, X_4), (X_2, X_5), and (X_3, X_6). We then sorted the first three indicators' values into ascending order. Then, we eliminated the last three indicators' values when the corresponding indicator of the pairs was ranked as the smallest p percentage. For MNAR, we sorted the last three indicators' values into ascending order and deleted their values if the values were ranked in the smallest p percentage.

In total, we examined 216 conditions for three missing data mechanisms (MCAR, MAR, and MNAR), four sample sizes (750, 1500 with 100 groups, 1500 with 50 groups, and 3000 samples), six ICCs (.050, .111, and .296 with the same between-group level but different within-group level parameters and .050, .111, and .296 with the same within-group level but different between-group level parameters), and three missing data proportions (0%, 15%, and 30%). For each condition, we conducted 1000 replications.

3.2 Data Analysis

In this study, six reliability measures were estimated and compared, single-level coefficient alpha α, within-group level coefficient alpha α_w, between-group level coefficient alpha α_B, single-level coefficient omega ω, within-group level coefficient omega ω_w, and between-group level coefficient omega ω_B. We used two missing data techniques, listwise deletion (LD) and full-information maximum likelihood (FIML) methods to handle the generated data sets with missing values.

To evaluate the reliability estimation regarding the simulated conditions, we used two criteria. To evaluate the accuracy of the estimation, the first criterion was the percentage bias calculated as

$$Percent\ Bias\ (\%) = \left[\frac{1}{1000} \sum_{r=1}^{1000} \frac{\left(\hat{\rho}_{XT_r} - \rho_{XT_r}\right)}{\rho_{XT_r}} \right] \times 100 \tag{10}$$

where $\hat{\rho}_{XT_r}$ was the estimated reliability measure in the r th replication, and ρ_{XT_r} was the population value under each simulated condition. We marked the condition showing bias above 15%, which indicated the condition would have a meaningful negative impact on the reliability estimation.

To evaluate the stability of the estimation, we calculated the second criterion, the convergence rate, as

$$Convergence\ Rate\ (\%) = \frac{(Number\ of\ Converged\ Models)}{1000} \times 100 \tag{11}$$

We defined the model as converged (a) when the model satisfied the good model-fit values (i.e., CFI > .90, TLI > .90, and RMSEA < .08) and (b) when there were no convergence error messages from the Mplus outputs. The larger value of the convergence rate indicated more stable results of the reliability estimation.

4 Results and Conclusions

4.1 Convergence Rates

In terms of convergence rates, because the single-level coefficient alpha and the within-group and between-group level coefficient alpha were estimated from the fully saturated model, the results showed a perfect model fit (i.e., 100% convergence rates) under all simulated conditions. For the coefficient omega estimates, the convergence rates were below 50% (39.8% under MAR, 35.5% under MNAR) when (a) we used the LD method, (b) the sample size was small (750 samples), and (c) the ICC value was large (.296) by fixing the same between-group level parameters, the two-level MCFA model to estimate the within-group and between-group coefficient omega ω_w and ω_B.

Our results of the lowest convergence rates (Figs. 2 and 3) extended the findings of Hancock and An (2020), because this simulation investigated the impact of multilevel related conditions (e.g., sample sizes at each level, ICCs) and missing data conditions (i.e., missing data mechanisms, missing data proportions).

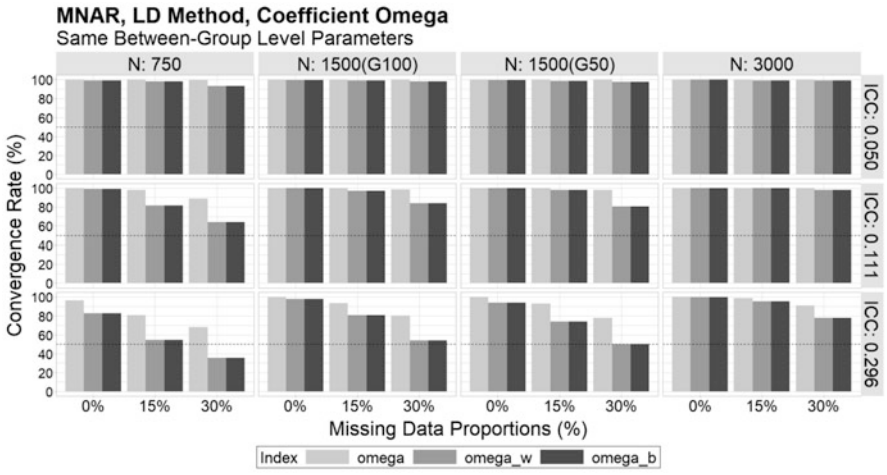


Fig. 2 Convergence rates of coefficient omega under MNAR using LD method with the low population values of the within-group reliabilities ($\alpha_w = .372$, $\omega_w = .376$)

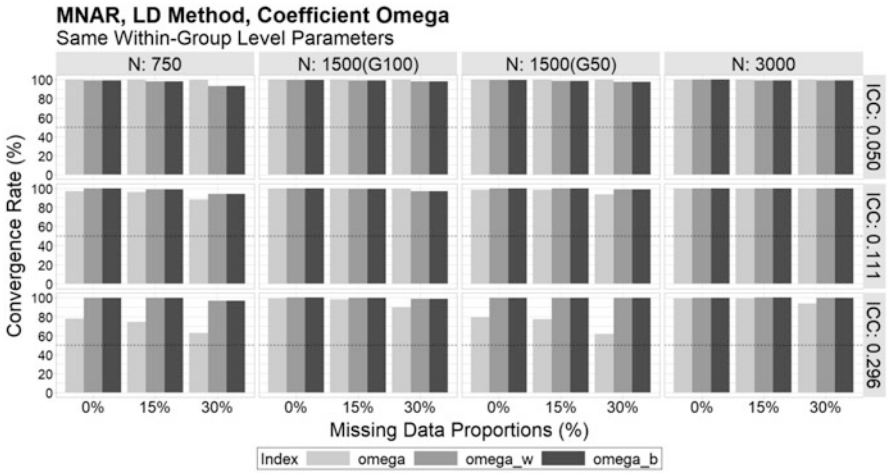


Fig. 3 Convergence rates of coefficient omega under MNAR using LD method with the low population values of the between-group level reliabilities ($\alpha_B = .372$, $\omega_B = .376$)

4.2 Percentage Bias

The single-level and within-group level reliability estimates reported the highest percentage bias under the conditions of (a) the small sample size (750 samples), (b) the large ICC value (.296), and (c) the low population values of the within-group reliabilities ($\alpha_w = .372$, $\omega_w = .376$). Meanwhile the between-group level reliability estimates provided the worst accuracy results under the conditions of (a) the small samples (750), (b) the small ICC value (.050), and (c) the low population values of the between-group level reliabilities ($\alpha_B = .372$, $\omega_B = .376$).

Regarding the missing data related factors, the conditions under (a) MNAR mechanism, (b) the more missing data proportions (30% missingness), and (c) the usage of LD method showed the worse performance of all six reliability measures, single-level and level-specific coefficient alpha and coefficient omega. The results of the percentage bias are corresponded with the results of convergence rates. Thus, it indicates that the simulation conditions related with the missing data have the meaningful impacts on the reliability estimation.

Further, under most simulated conditions in this study, the percentage bias of the coefficient alpha was higher than the percentage bias of the coefficient omega. In the data generation step of the simulation study, we set the population values of the factor loadings and the error variances varied, because it is a case of the congeneric measurement model.

The literatures pointed out how the violation of the tau-equivalent assumptions could cause the overestimation issue of single-level coefficient alpha (Deng & Chan, 2017; Green & Yang, 2009; Raykov & Marcoulides, 2006), and the level-specific coefficient alpha (Geldhof et al., 2014). The results of this study not only supported

the findings of the previous studies, but also showed the importance of checking the assumptions underlying the test (e.g., tau-equivalent, congeneric) under the multilevel data structure.

5 Discussions

To report reliability is one of the conventions in social science studies because reliability ensures the test results' consistency. However, Cortina (1993) and Sijtsma (2009) pointed out that the traditional measures of reliability measures, such as Cronbach alpha or coefficient omega, would not perform well under complicated test administering situations.

For example, if the researcher repeatedly administers a test to the students (by semesters or by years), the test results would have a hierarchy. Also, if the researcher is interested in both the students' and schools' features simultaneously using a questionnaire, then the collected data would be analyzed by the multilevel model. In these cases, when reporting the reliabilities of the test or questionnaire, the traditional reliabilities might be biased because they cannot handle the data hierarchy. Figures 4, 5, 6, and 7 show that when the ICC is large (.296), indicating the multilevel model performs better than a single-level one, the between-group level reliabilities outperformed the single-level reliabilities. Under the high ICC conditions (i.e., above .3), the between-group level coefficient alpha and coefficient omega showed lower percentage bias. Thus, the results of this study support the necessity of level-specific reliabilities, as suggested by Geldhof et al. (2014) and Hancock and An (2020).

In addition to the multilevel data structure, the missing-data-related condition would be the critical factor of reliability estimation. When the examinee skips answering some items of the test or when the researcher mistakenly omits responses to the questionnaire, missing data could occur. The existence of missing data causes bias in the reliability measures because of the loss of the complete test information. Even though the impacts of missing data on reliabilities were examined by previous studies (Deng & Chan, 2017; Edwards et al., 2021; Enders, 2003, 2004), the authors did not consider the multilevel data structure. The findings of this study showed the impacts of missing-data-related conditions on reliability estimation, including the missing data mechanisms, missing data proportions, and missing data techniques. Specifically, the reliability estimates showed the most biased results when the missing data were generated under MNAR, over 30%, and using the LD method. It implies that when we report the reliability, we should consider these three conditions if the data set has missingness.

This study could be extended. First, more missing data techniques, such as the data imputation methods, can be employed and compared. Because it replaces missing information with the most plausible values instead of eliminating missingness, the mean imputation method showed better performance than LD methods in single-level reliability estimation (Enders, 2003, 2004). Thus, we could investigate

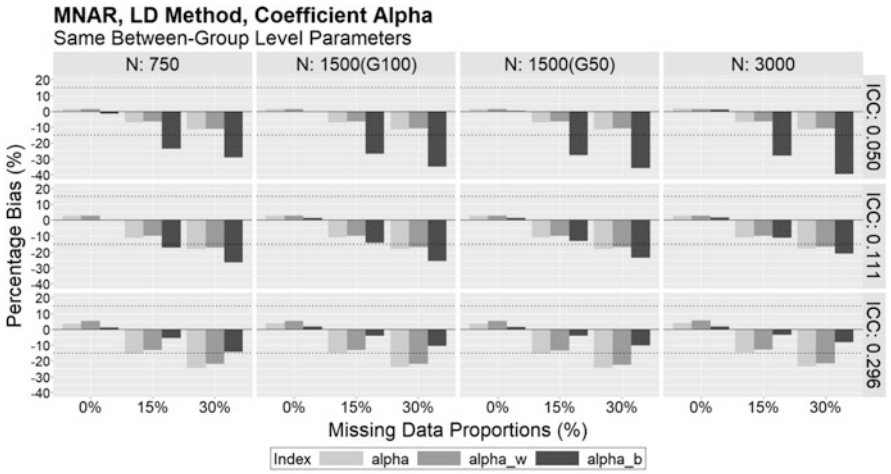


Fig. 4 Percentage bias of coefficient alpha under MNAR using LD method with the low population values of the within-group reliabilities ($\alpha_w = .372$, $\omega_w = .376$)

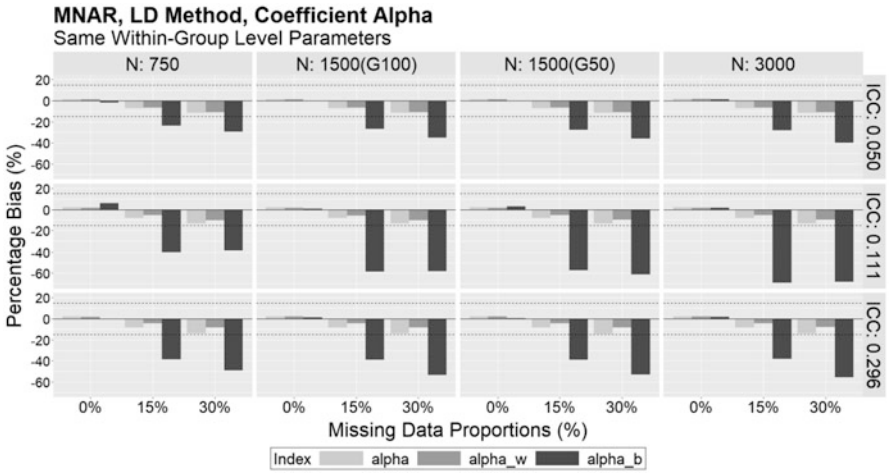


Fig. 5 Percentage bias of coefficient alpha under MNAR using LD method with the low population values of the between-group level reliabilities ($\alpha_B = .372$, $\omega_B = .376$)

the data imputation methods to handle the missing data in level-specific reliability estimation.

Second, in addition to the sample sizes and the ICCs examined, the effects of other data features, such as the number of items (i.e., the test length), the number of the answer categories (i.e., the point scale of the items), the complexity of the model, or the misspecification of the model could be considered in future research.

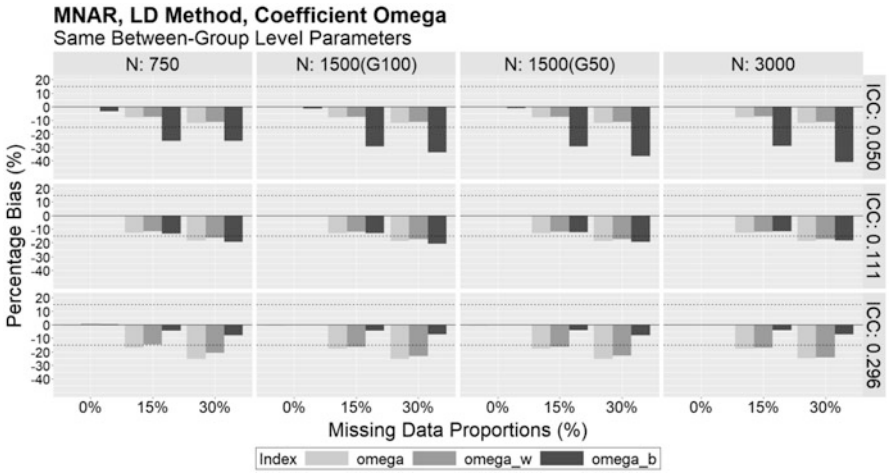


Fig. 6 Percentage bias of coefficient omega under MNAR using LD method with the low population values of the within-group reliabilities ($\alpha_w = .372$, $\omega_w = .376$)

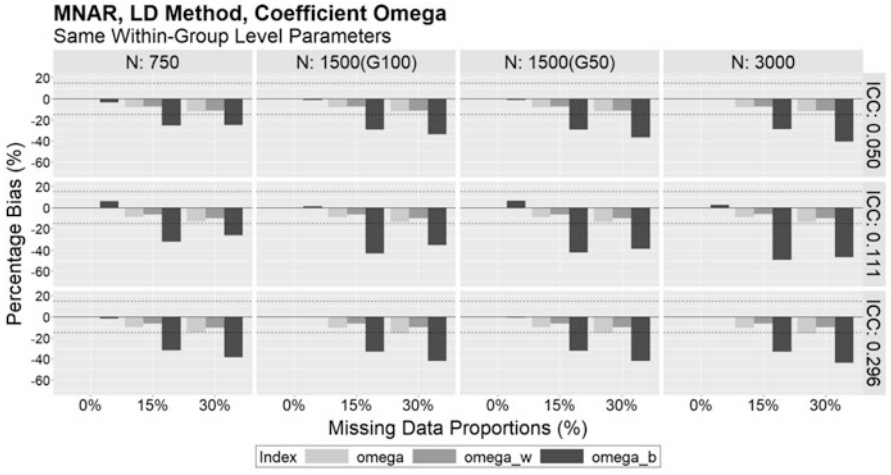


Fig. 7 Percentage bias of coefficient alpha under MNAR using LD method with the low population values of the between-group level reliabilities ($\alpha_B = .372$, $\omega_B = .376$)

References

Bonito, J. A., Ruppel, E. K., & Keyton, J. (2012). Reliability estimates for multilevel designs in group research. *Small Group Research, 43*(4), 443–467.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101*(1), 147–158.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Deng, L., & Chan, W. (2017). Testing the difference between reliability coefficients alpha and omega. *Educational and Psychological Measurement, 77*(2), 185–203.

Edwards, A. A., Joyner, K. J., & Schatschneider, C. (2021). A simulation study on the performance of different reliability estimation methods. *Educational and Psychological Measurement, 81*(6), 1089–1117.

Enders, C. K. (2003). Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data. *Psychological Methods, 8*(3), 322–337.

Enders, C. K. (2004). The impact of missing data on sample reliability estimates: Implications for reliability reporting practices. *Educational and Psychological Measurement, 64*(3), 419–436.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*(1), 72–91.

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika, 74*(1), 121–135.

Hancock, G. R., & An, J. (2020). A closed-form alternative for estimating ω reliability under unidimensionality. *Measurement: Interdisciplinary Research and Perspectives, 18*(1), 1–14.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*(2), 109–133.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley.

McDonald, R. P. (1978). Generalizability in factorable domains: "Domain validity and generalizability". *Educational and Psychological Measurement, 38*(1), 75–79.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*(3), 376–398.

Raykov, T. (2009). Evaluation of scale reliability for unidimensional measures using latent variable modeling. *Measurement and Evaluation in Counseling and Development, 42*(3), 223–232.

Raykov, T., & Marcoulides, G. A. (2006). On multilevel model reliability estimation from the perspective of structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 13*(1), 130–141.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107–120.

New Flexible Item Response Models for Dichotomous Responses with Applications

Jessica Suzana Barragan Alves and Jorge Luis Bazán

Abstract Some asymmetric Item Characteristic Curves (ICC)s have already been introduced in the IRT literature. These proposals include a new item parameter associated with the item complexity which explains the asymmetry in the ICC. Although the importance of proposing new models that have asymmetric ICC in IRT is already known, the relationship between these models and unbalanced binary responses in testing data in real applications has not been explored. In this work we propose new asymmetric IRT models that have an asymmetric ICC as their main feature. A special case of these models is the cloglog IRT model. Bayesian estimation of the proposed models is discussed and one application in educational data illustrates the benefits of the new ICC when we compare our IRT models with other IRT models proposed in the literature.

Keywords Asymmetric ICC · Bayesian estimation · LPE · Gumbel distribution

1 Introduction

Item Response Theory (IRT) is concerned with modeling the relationship between the probability of an individual selecting a certain response to an item and the individual's latent traits (characteristics of the individual that cannot be directly observed or measured) (Hambleton et al., 1991). Specifically, in this work we are interested in working with dichotomous item responses, modeling the probability of

These authors contributed equally to this work.

J. S. B. Alves (✉) · J. L. Bazán
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo e Departamento de Estatística, Universidade Federal de São Carlos, USP/UFSCAR, São Carlos, SP, Brasil

Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - USP, São Carlos, SP, Brasil
e-mail: jessicasbarragan@usp.br; jlbazan@icmc.usp.br

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_27

selecting the correct response, namely, p_{ij}, as $p_{ij} = F(m_{ij})$, where F is called the item's characteristic curve (ICC), $m_{ij} = a_j(\theta_i - b_j)$, for $i = 1, \ldots, n$ individuals with, $j = 1, \ldots, k$ items, where a_j and b_j are parameters associated with the items, and θ_i is the latent variable associated with the individuals. Parameters a_j and b_j are called discrimination and difficulty parameters, respectively.

Traditional item response theory (IRT) models, such as the two- and three-parameter logistic models or normal ogive, have a symmetric ICC. That is, the behavior of ICCs observed to the right of the inflection point is a mirror image of what happens to the left of the inflection point (De Ayala, 2013).

Some asymmetric ICCs have already been introduced in IRT literature. Samejima (2000) introduced an exponent parameter in its Logistic Positive Exponent Model (LPE). This new parameter included in the exponent was defined as the item complexity and produced asymmetry in the ICC. J.L. Bazán, Bolfarine, and Branco (2006) in turn proposed a family of asymmetric ICCs called skew-normal IRT models, while Bolfarine and Bazán (2010) proposed a new model that is a reflection of Samejima's LPE model and is called the Reflection Logistic Positive Exponent Model (RLPE) (Samejima, 2000). Recently, Zhang et al. (2022) proposed a family of IRT models with generalized logit links, which include the traditional logistic and normal ogive models as special cases.

Although the importance of proposing new models that have asymmetric ICC in IRT is already well-established, as far as we know, generalizations of the *cloglog* ICC have not been introduced in the literature. Motivated by the proposal of new links in the context of binary regression in Alves et al. (2022), we will propose some of the links in the context of item response theory. All models can be considered asymmetric IRT models and have an asymmetric ICC as their main feature. Estimation will be developed using a Bayesian approach, specifically the NUTS algorithm of the Stan software, which can be used to simulate from posterior distributions of item parameters and latent variables (Stan Development Team, 2020). The NUTS algorithm is an extension of the Monte Carlo algorithm proposed by Hoffman and Gelman (2014). This algorithm allows the Markov chain to explore the objective distribution much more efficiently than other widely known MCMC methods such as Metropolis Hastings and Gibbs sampling, using Hamiltonian dynamics instead of a probability distribution allowing the Markov chains to converge quickly (Neal, 2011).

The main objective is to provide a clear presentation of Bayesian estimation via MCMC for the considered asymmetric IRT models.

The paper is organized as follows: Sect. 2 will describe traditional IRT models, in Sect. 3 we present flexible dichotomous models of the Gumbel distribution, which can be considered as ICCs for IRT models. In Sect. 4, the Bayesian analysis, comparison criteria and residual analysis used in the application are presented. Section 5 exemplifies the methodology through an application.

2 Traditional IRT Models

Let Y_{ij} be the random variable that denotes the i-th individual's response to item j, where $i = 1, \ldots, n$ and $j = 1, \ldots, k$. The i-th individual's response pattern is written as $Y_i = (Y_{i1}, \ldots, Y_{ik})$. When items are scored dichotomically (correct, incorrect), the observed data can assume the values $Y_{ij} = 1$, for a correct answer, and $Y_{ij} = 0$, otherwise. It is also assumed that of the event $Y_{ij} = 1$ (correct answer), is denoted by p_{ij}, and can be written as

$$p_{ij} = P\left[Y_{ij} = 1 \mid \theta_i, a_j, b_j\right] = F(m_{ij}), \tag{1}$$

where F is called the ICC, and

$$m_{ij} = a_j(\theta_i - b_j), \qquad i = 1, \ldots, n, \qquad j = 1, \ldots, k. \tag{2}$$

is a latent linear predictor where a_j and b_j are parameters associated with the items, and θ_i is the latent variable associated with the latent ability or trait of the i-th individual. The random variables Y_{ij} associated with the items are conditionally independent, given θ_i. As previously indicated, parameters a_j and b_j are called discrimination and difficulty parameters, respectively.

The first binary IRT model was introduced by Lord (1952). Equation (1) was the cumulative distribution function (cdf) of the standard normal distribution. In addition, Birnbaum (1968) considered the default logistic distribution cdf. These models are often referred to as the normal ogive IRT model and the logistic IRT model, respectively, denoted here as the 2P and 2L IRT models. When $a_j = 1$, for $j = 1, \ldots, k$ in Eq. (2), we obtain the 1P and 1L models considering only the item's difficulty parameters. Furthermore, we can consider the models with three parameters as 3P and 3L, proposed by Sahu (2002) which are obtained by considering $p_{ij} = c_j + (1 - c_j)F(m_{ij})$ in Eq. (1), where c_j is the guessing parameter, indicating that the probability of a correct answer is greater than zero.

The traditional IRT models presented above have a symmetric ICC. Recently, the cloglog link was used by Robitzsch (2022), as an asymmetric link. However, it has been noted by Samejima (2000), J.L. Bazán et al. (2006) as well as Bolfarine and Bazán (2010) that asymmetric ICCs can be incorporated using a new item parameter that controls the shape of the curve.

The Logistic Positive Exponent Model (LPE) was proposed by Samejima (2000). The reflection of the Logistic Positive Exponent or Reverse Logistic Positive Exponent Model (RLPE) was formulated by Bolfarine and Bazán (2010). These models assume that: $Y_{ij} \mid \theta_i, a_j, b_j \sim Bernoulli(p_{ij})$, where $p_{ij} = P(Y_{ij} = 1 \mid \theta_i, a_j, b_j) = F_{\lambda_j}(m_{ij})$, $i = 1, \ldots, n$ and $j = 1, \ldots, k$. where $\lambda_j > 0$ is the shape parameter; and F_{λ_j} is a cumulative distribution function indexed by λ_j. The LPE and RLPE correspond to c.d.f. $F(m_{ij}) = (1 + \exp(-m_{ij}))^{-\lambda_j}$ and $F(m_{ij}) = 1 - (1 + \exp(m_{ij}))^{-\lambda_j}$ respectively. Note that when $\lambda_j = 0$ in the LPE model we have a special case of the IRT 2L model.

As in binary regression, this asymmetry is necessary in situations where responses with low or high proportions of ones are observed. Some authors have noted that incorrect specification of the link function can result in considerable bias in the mean response estimates (Czado & Santner, 1992). As emphasized in Chen et al. (2001), symmetric links do not always provide a good fit for some datasets. Alves et al. (2022) emphasizes that some generalizations of the cloglog link turn out to be interesting when we have imbalanced data in the context of binary regression. This way, we will extend the links presented in Alves et al. (2022) in the context of IRT models.

3 Flexible Dichotomous IRT Models

In the following section, we propose new dichotomous IRT models by modifying Eq. (1). Our proposal starts with the construction of the power and reverse power distributions proposed by J. Bazán et al. (2017) and Lemonte and Bazán (2018):

$$p_{ij} = F_{P-RG}(m_{ij}) = \left[1 - \exp\left\{-\exp(m_{ij})\right\}\right]^{\delta_j}, \tag{3}$$

and

$$p_{ij} = F_{RP-RG}(m_{ij}) = 1 - \left[1 - \exp\left\{-\exp(-m_{ij})\right\}\right]^{\delta_j}, \tag{4}$$

with $\delta_j > 0$. Moreover, in (3) and (4) a convenient reparameterization is given by $\delta_j = \exp(\lambda_j)$ for $\lambda_j \in \mathbb{R}^n$. When $\lambda_j = 0, \forall j = 1, \ldots, k$ we will have the Gumbel IRT model and the Reverse Gumbel IRT model as a particular cases, respectively. We will call these models *standard Power Reverse Gumbel* (P-RG) IRT model and *standard Reverse Power Reverse Gumbel* (RP-RG) IRT model, respectively, because we are using the corresponding cdf of these distributions.

Furthermore, we will also consider the process based on building Transmuted Skew distributions by following Shaw and Buckley (2009). The resulting models are given by:

$$p_{ij} = F_{TS-G}(m_{ij}) = \exp\left\{-\exp(-m_{ij})\right\}\left(1 + \delta_j\left[1 - \exp\left\{-\exp(-m_{ij})\right\}\right]\right), \tag{5}$$

and

$$p_{ij} = F_{TS-RG}(m_{ij}) = \left[1 - \exp\left\{-\exp(m_{ij})\right\}\right]\left(1 + \delta_j \exp\left\{-\exp(m_{ij})\right\}\right), \tag{6}$$

where $\mid \delta_j \mid < 1$. We considered a reparameterization for (5) and (6) given by $\delta_j = \frac{e^{\lambda_j}-1}{e^{\lambda_j}+1}$ for $\lambda_j \in \mathbb{R}^n$. Note that if $\lambda_j = 0$, we will have the IRT Gumbel and Reverse Gumbel models, respectively, as individual cases. These models are called the *Transmuted Skew Gumbel* (TS-G) IRT model and *Transmuted Skew Reverse*

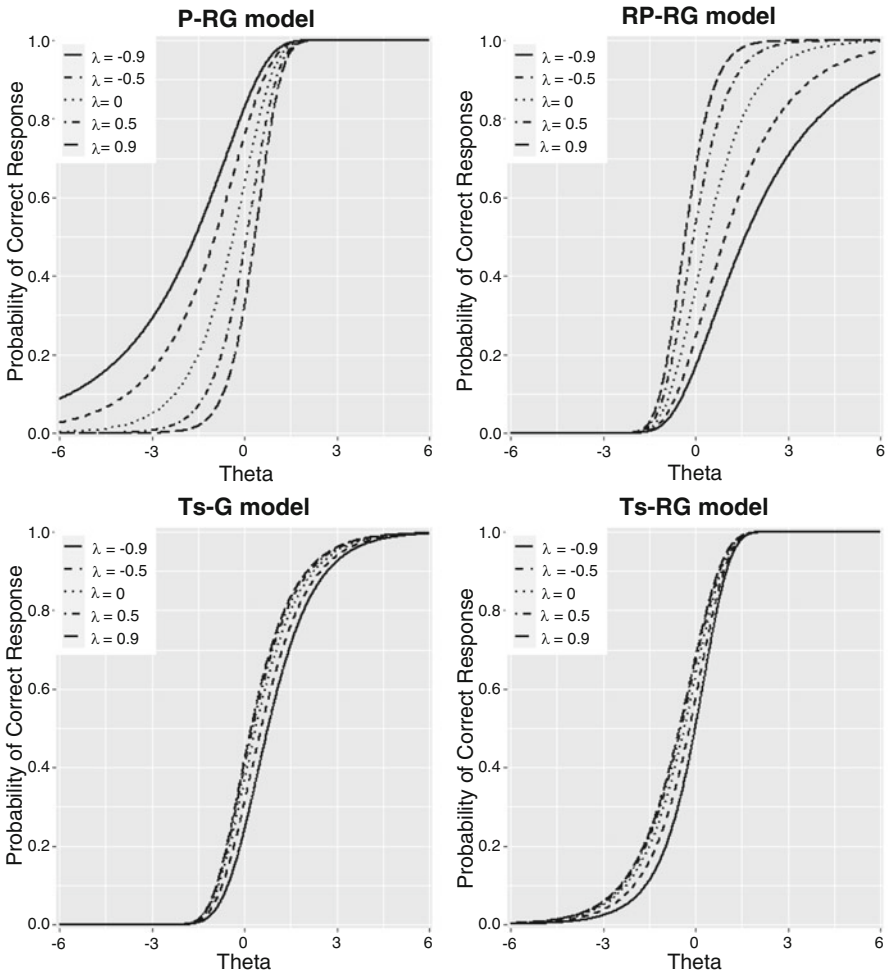


Fig. 1 Probability curves for $\lambda = -0.9, -0.5, 0, 0.5, 0.9$ in P-RG, RP-RG, TS-G and TS-RG models considering different ranges for θ and $a = 1$ and $b = 0$

Gumbel (TS-RG) IRT model, respectively, because we are using the corresponding cdf of these distributions.

Figure 1 represents the ICCs where the probability of success is given as a function of θ, a (discrimination parameter) and b (difficulty parameter), which were set to 1 and 0, respectively. With this we can say that if an item has a more accentuated ICC then it can be considered as having a high power of discrimination.

For $0 < \lambda_j < 1$, the ICC of the P-RG IRT model is below the ICC of the model with the Gumbel link function, and if $\lambda_j < 0$ the ICCs of both models are above the ICC of the model with the Gumbel link function. The opposite happens when we consider the RP-RG IRT model, that is, if $0 < \lambda_j < 1$ we have the ICC of the

RP-RG IRT model above the model with the Reverse Gumbel link function, and if $0 < \lambda_j$, we have the RP-RG IRT ICC model below the Reverse Gumbel IRT model.

The TS-G and TS-RG IRT models behave similarly to λ_j, that is, if we have $0 < \lambda_j < 1$ the ICC of these models will be above the ICC of the model with Gumbel link, and if $\lambda_j < 0$ the ICC will be below the ICC of the Gumbel IRT model.

The λ_j parameter in the proposed IRT models is associated with the asymmetric form of the ICC, or equivalently, with the asymmetry considered in the latent error of the regression of the latent auxiliary variable in relation to the underlying latent trait of the correct or incorrect answer of the item. This extra parameter is interpreted in two ways in the literature, as a penalty parameter (Bolfarine & Bazán, 2010) or as an acceleration parameter by Samejima (2000). Both terms will be used in this work without distinction, because they are part of a complexity parameter (see, Bolt & Liao, 2022).

4 Inference

Likelihood Function For Expressions 3 to 6 we use $F_{\lambda_j}(m_{ij})$ as general notation. For the flexible *cloglog* models indexed by λ_j, the likelihood function is given by

$$L(\boldsymbol{\beta}, \theta \mid \boldsymbol{y}, \boldsymbol{X}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \left\{F_{\lambda_j}(m_{ij})\right\}^{y_{ij}} \left\{1 - F_{\lambda_j}(m_{ij})\right\}^{1-y_{ij}}, \tag{7}$$

where $\beta = (\boldsymbol{a}^T, \boldsymbol{b}^T)$, $\boldsymbol{a} = (a_1, \ldots, a_n)^T$, $\boldsymbol{b} = (b_1, \ldots, b_n)^T$, m_{ij} is the latent linear predictor in Eq. 2 for $i = 1, \ldots, n$ and $j = 1, \ldots, k$, and F_{λ_j} is the cdf P-RG, RP-RG, TS-G and TS-RG in Eq. (3–6), indexed with λ_j and evaluated at m_{ij}.

Bayesian Approach In this work, we specifically consider Bayesian estimation with Markov Chain Monte Carlo methods that facilitate the efficient sampling of the posterior marginal distribution of the parameters of interest. This choice for Bayesian estimation was made considering the works of Swaminathan et al. (2003) who demonstrated that the accurate estimation of the parameters of items in small samples is obtained by using the Bayesian approach. To do this, we will use the `rstan` package in R Development Core Team (2009) software that uses the No-U-Turn Sampler (NUTS) algorithm (Hoffman & Gelman, 2014). Luo and Jiao (2018) demonstrated in a comparison study with other BUGS software that STAN is considerably faster in estimating IRT models.

Prior Specification Prior specification is an important aspect of Bayesian analysis especially in the case of small sample size (Bolfarine & Bazán, 2010). Choosing an adequate prior distribution in the latent trait solves particular identification problems (Albert & Ghosh, 2000). In the IRT literature there is a consensus regarding the specification of the prior for the latent trait $\theta_i \sim N(0, 1)$, for $i = 1, \ldots, n$.

There is no consensus in the literature regarding a_j and b_j parameters of the items, (see, Albert & Ghosh, 2000; Congdon, 2007; Johnson & Albert, 1999; Patz & Junker, 1999; Sahu, 2002; Spiegelhalter et al., 1996). Therefore, we choose to use independent and common priors for a, b, and λ and let such correlations be data dependent as in Bolfarine and Bazán (2010). That is, the prior we consider can be written as

$$\pi(\theta, \boldsymbol{a}, \boldsymbol{b}, \lambda) = \prod_{i=1}^{n} \Phi(\theta_i) \prod_{j=1}^{k} \pi_1(a_j)\pi_2(b_j)\pi_3(\lambda_j) \tag{8}$$

where $\Phi(.)$ is the pdf of the standard normal distribution and $\pi_1(.)$, $\pi_2(.)$, $\pi_3(.)$ are the prior pdf for parameters a_j, b_j, and λ_j, respectively. This way, $a_j \sim LN(0, 1)$ and $b_j \sim N(0, 1)$ (see, Bolfarine and Bazán (2010)).

A prior considered for λ_j was $U(-2, 2)$ based on Alves et al. (2022) with binary regression data.

Model Comparison Criteria Several methodologies exist to compare alternative models in a Bayesian framework. In this work, we consider the Watanabe-Akaike information criterion (WAIC) based on the complexity parameter (p_{WAIC}), proposed by Watanabe and Opper (2010), and defined as: $WAIC = -2lppd + 2p_D$ where $lppd$ is the log pointwise predictive density and p_D is the effective number of parameters. Another measure called 'leave-one-out cross-validation' (LOO) proposed by Geisser and Eddy (1979) was used. Finally, Deviance Information Criterion (DIC) using the definition found in Brooks (2002): $DIC = \overline{D(\boldsymbol{\theta})} + 0.5var(D(\boldsymbol{\theta}))$, where $\overline{D(\boldsymbol{\theta})}$ is the posterior mean deviance was used for model comparison. For all the above criteria, smaller values indicate better fit.

5 Application

An application was used to illustrate the Bayesian approaches developed in this work for IRT models with binary responses. We used data from Bolfarine and Bazán (2010), which corresponds to a math test for fourth-grade students in Peruvian rural primary schools. There are 974 students responses to 18 items that qualify as binary (correct or incorrect) responses. These data are unbalanced and have proportions of ones for each item given by (0.72, 0.61, 0.43, 0.37, 0.50, 0.08, 0.65, 0.27, 0.53, 0.80, 0.48, 0.67, 0.47, 0.57, 0.30, 0.28, 0.12, 0.42).

We will show a study on the fit of the parametric IRT models discussed earlier in Sect. 3 using the mathematical test data. Logistic IRT models with two parameters (2L), LPE and RLPE models proposed by Bolfarine and Bazán (2010) are considered in our comparison. In all cases, the prior distributions used are $\theta \sim N(0, 1), b \sim N(0, 1), a \sim LN(0, 1)$. For λ, we consider a $U(-2, 2)$ prior. The stan code used in this application for the RP-RG model is available at Appendix.

Table 1 Model comparison criteria for Math-test data

	DIC	WAIC	LOO	$\hat{R}_a$	$\hat{R}_b$	$\hat{R}_\lambda$	$\hat{R}_\theta$
2L	16,332.96	17,754.66	17,769.34	1.00	1.00	–	1.00
LPE	16,281.30	17,731.34	17,753.75	1.00	1.00	1.00	1.00
RLPE	16,262.84	17,713.35	17,737.99	1.01	1.00	1.01	1.00
P-RG	16,230.57	17,723.74	17,744.28	1.12	1.14	1.14	1.11
RP-RG	**16,258.68**	**17,694.58**	**17,726.59**	**1.00**	**1.00**	**1.00**	**1.00**
TS-G	16,273.34	17,701.97	17,726.57	1.05	1.06	1.06	1.02
TS-RG	16,371.22	17,795.55	17,817.49	1.11	1.15	1.15	1.03

Table 2 Item parameters for alternative IRT models for item 14, item 2, and item 11 in Math data

		Items parameters					
		Discrimination		Difficulty		Acceleration	
Items	Models	a	$\hat{R}_a$	b	$\hat{R}_b$	λ	$\hat{R}_\lambda$
Item 14 (Asymmetric)	RPLE	2.69	1.00	−1.51	1.01	−1.41	1.01
	RP-RG	1.95	1.00	−1.41	1.00	−0.98	1.00
Item 02 (some degree of asymmetry)	RPLE	1.05	1.00	0.48	1.00	0.69	1.00
	RP-RG	0.54	1.00	0.36	1.00	1.00	1.00
Item 11 (Asymmetry is not evident)	RPLE	1.98	1.00	0.15	1.00	0.01	1.00
	RP-RG	0.95	1.00	0.33	1.00	0.67	1.00

In this application, we apply the estimation algorithm using R. With the `rstan` R language package, the following MCMC parameters were considered: 20,000 iterations with 10,000 burn-in iterations, along with a thinning interval of 10 iterations to achieve convergence using three MCMC chains. We consider the convergence check based on the potential reduction statistic ($\hat{R}$) (Gelman & Rubin, 1992). To compare the models, we used the DIC, WAIC, and LOO selection criteria discussed in Sect. 4. These values are shown in Table 1.

We can observe in Table 1 that the best proposed model considering the criteria studied is the RP-RG model. Note that for all parameters in the proposed models the $\hat{R}$ was around 1, which indicates that no convergence problems were detected.

We now compare the IRT RP-RG and RLPE models. Thus, we will continue the analysis describing 3 items: Item 14, Item 2 and Item 11, which are the same items analyzed by Bolfarine and Bazán (2010). For the RLPE model, item 14 is considered an asymmetric item, item 2 has some degree of asymmetry; and item 11 has no asymmetry (for more details see Bolfarine & Bazán, 2010).

Table 2 shows the item parameters for the alternative IRT model that had the best performance when compared to the RLPE. It is worth mentioning that the results obtained in this paper consider $\lambda \sim U(-2, 2)$ and a change of variable $\delta = exp(\lambda)$, in this document we will call this the indirect prior, which is different from those used in the article by Bolfarine and Bazán (2010), where $\lambda \sim Gamma(0.25, 0.25)$ is used. The $U(-2, 2)$ prior was chosen because Alves et al. (2022) shows that better

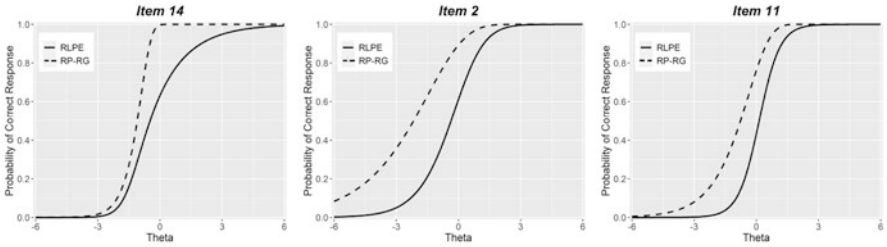


Fig. 2 Item characteristic curves (ICCs) for Items 14, 2, and 11 under the RP-RG and RLPE IRT models in the math data

results were obtained with said prior than those using the direct prior for binomial data.

When comparing the RP-RG and RLPE IRT models in terms of the λ parameter, we observed different magnitudes in the asymmetric case, but with equal signs. In the asymmetric case the magnitudes are similar. We also observed $\lambda = 0.67$ in Item 11 (case in which asymmetry is not evident) for the RP-RG IRT model. This is expected for this model, as we know that it does not detect symmetric items. Regarding the difficulty and discrimination parameters, we observed equal signs. As for the magnitude, we observed a difference between the RP-RG and RLPE models for the discrimination parameter, but this magnitude is similar for the difficulty parameters. Note that for all parameters in the proposed models we find $\hat{R} \approx 1$.

We can observe in Fig. 2 that for item 14, the ICC of the RP-RG model shows that if the individual has no skill or this skill is small ($\theta < -3$), the probability of correctly answering the item is practically zero, but if he has some skill level ($-2.5 < \theta < 0$) the probability of correctly answering the item increases significantly. And if this skill is greater than 0, the probability of correctly answering the item is practically 1. Note that the RP-RG and the RLPE show that this item is asymmetric.

For item 2, note the probability of correctly answering the item even if the individual has an extremely low ability ($\theta < -6$) is non-zero for the RP-RG model. Now if the skill is greater than 0 the probability of answering correctly is very near to 1. This means that this item is not an extremely complex item. Note that in item 11 the results for lambda was 0.01 for the RPLE model, which is a particular case where the curve is symmetric. As for the RP-RG model, we have $\lambda = 0.67$ which also approaches zero, so the curve is close to the Reverse Gumbel model. We can thus say that item 11 is an item in which the individual does not need to have sequential sub-processes to select a correct answer and therefore it is a symmetric item.

Figure 3 shows the relationship between the discrimination, difficulty and acceleration parameters for the IRT RP-RG model. We can see that there is some relationship between the discrimination parameters and the λ parameter, that is, as the value of the discrimination parameter(a) increases, the value of the precision parameter(λ) decreases. The other parameters are apparently unrelated.

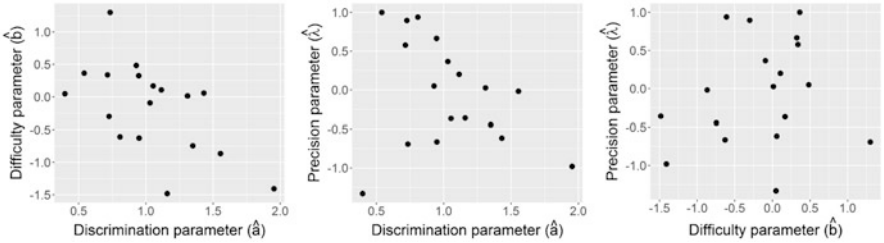


Fig. 3 Relationship between the discrimination, difficulty and acceleration parameters, for the IRT RP-RG model

6 Final Comments

In this article, we propose generalizations of link functions based on the *cloglog* ICC as an alternative to traditional Item Response Theory models. Note that cloglog was used by Robitzsch (2022) and is a special case of the model proposed there. We developed a Bayesian estimation procedure for item response theory models. In this study, we observed the importance of using the flexible ICC regarding the asymmetry of the latent variable.

In the application referring to the Mathematics Test data, the RP-RG IRT model obtained a better fit than the IRT models proposed by Bolfarine and Bazán (2010), thus being an alternative to work with unbalanced dichotomous data in IRT.

Appendix: RP-RG Stan Code

```
data {
 int<lower=0> n;
 int<lower=0> p;
 int<lower=0,upper=1> Y[n,p];
 }
 parameters {
 vector[n] theta;
 vector[p] b;
 vector<lower=0>[p] a;
 vector[p] lambda;
}
transformed parameters{
vector[p] m[n];
vector[p] pp[n];
vector<lower=0,upper=1>[p] prob[n];
vector <lower=0>[p] delta;
```

```
for(i in 1:n){
 for (j in 1:p){
   delta[j] = exp(lambda[j]);
     m[i,j] = a[j]*(theta[i]-b[j]);
    pp[i,j] = exp(-exp(-m[i,j]));
 prob[i,j] = 1 - pow((1-pp[i,j]), delta[j]);
}
}
}
model {
    theta ~ normal(0,1);
     b ~ normal(0,1);
     a ~ lognormal(0,1);
    lambda ~ uniform(-2,2);
  for(i in 1:n){
   for (j in 1:p){
    Y[i,j] ~ bernoulli(prob[i,j]);
   }
  }
 }
 generated quantities {
 vector[p] loglik_y[n];
 vector[p] Y_rep[n];
  for (i in 1: n){
   for (j in 1: p){
     loglik_y[i,j] = bernoulli_lpmf(Y[i,j]|prob[i,j]);
        Y_rep[i,j] = bernoulli_rng(prob[i,j]);
   }
  }
 }
```

References

Albert, J., & Ghosh, M. (2000). Item response modeling. In D. K. Dey, S. K. Ghosh, & B. K. Mallick (Eds.), Generalized linear models: A Bayesian perspective (pp. 174–193).

Alves, J. S., Bazán, J. L., & Arellano-Valle, R. B. (2022). Flexible cloglog links for binomial regression models as an alternative for imbalanced medical data. *Biometrical Journal, 65*(3), 2100325.

Bazán, J., Torres-Avilés, F., Suzuki, A. K., & Louzada, F. (2017). Power and reversal power links for binary regressions: An application for motor insurance policyholders. *Applied Stochastic Models in Business and Industry, 33*(1), 22–34.

Bazán, J. L., Bolfarine, H., Branco, M.D. (2006). A skew item response model. *Bayesian Analysis, 1*(4), 861–892.

Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. In *Statistical theories of mental test scores*. Addison-Wesley.

Bolfarine, H., & Bazán, J. L. (2010). Bayesian estimation of the logistic positive exponent IRT model. *Journal of Educational and Behavioral Statistics, 35*(6), 693–713.

Bolt, D. M., & Liao, X. (2022). Item complexity: A neglected psychometric feature of test items? *Psychometrika, 87*(4), 1195–1213.

Brooks, S. (2002). Discussion of the paper Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, 64*, 616–618.

Chen, M.-H., Dey, D. K., Shao, Q.-M. (2001). Bayesian analysis of binary data using skewed logit models. *Calcutta Statistical Association Bulletin, 51*(1–2), 11–30.

Congdon, P. (2007). *Bayesian statistical modelling* (Vol. 704). John Wiley & Sons.

Czado, C., & Santner, T. J. (1992). The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference, 33*(2), 213–231.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association, 74*(365), 153–160.

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Hoffman, M. D., & Gelman, A. (2014). The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning and Research, 15*(1), 1593–1623.

Johnson, V. E., & Albert, J. H. (1999). Regression models for ordinal data. In *Ordinal data modeling* (pp. 126–157). Springer.

Lemonte, A. J., & Bazán, J. L. (2018). New links for binary regression: an application to coca cultivation in Peru. *Test, 27*(3), 597–617.

Lord, F. (1952). *A theory of test scores*. *Psychometric monographs*. Psychometric Corporation.

Luo, Y., & Jiao, H. (2018). Using the Stan program for Bayesian item response theory. *Educational and Psychological Measurement, 78*(3), 384–408.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo, 2*(11), 116–162.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain monte carlo methods for item response models. *Journal of educational and behavioral Statistics, 24*(2), 146–178.

Robitzsch, A. (2022). On the choice of the item response model for scaling pisa data: Model selection based on information criteria and quantifying model uncertainty. *Entropy, 24*(6), 760.

R Development Core Team (2009). R: A language and environment for statistical computing [Computer software manual]. Retrieved from http://www.R-project.org (ISBN 3-900051-07-0)

Sahu, S.K. (2002). Bayesian estimation and model choice in item response models. *Journal of Statistical Computation and Simulation, 72*(3), 217–232.

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

Shaw, W. T., & Buckley, I. R. C. (2009, January). *The alchemy of probability distributions: beyond Gram-Charlier expansions, and a skew-kurtoticnormal distribution from a rank transmutation map*. ArXiv e-prints. https://arxiv.org/abs/0901.0434 [q-fin.ST]

Spiegelhalter, D., Thomas, A., Best, N., & Gilks, W. (1996). *Bugs 0.5 examples* (vol. 1 version I). Cambrigde, UK: University of Cambride.

Stan Development Team. (2020). *RStan: the R interface to Stan*. Retrieved from http://mc-stan.org/ (R package version 2.21.2)

Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. M. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement, 27*(1), 27–51.

Watanabe, S., & Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research, 11*(12), 3571–3594.

Zhang, J., Zhang, Y.-Y., Tao, J., & Chen, M.-H. (2022). Bayesian item response theory models with flexible generalized logit links. *Applied Psychological Measurement, 46*, 382, 01466216221089343.

Estimating Individual Dynamic Factor Models Using a Regularized Hybrid Unified Structural Equation Modeling with Latent Variable

Ai Ye and Kenneth A. Bollen

Abstract There has been an increasing call to model multivariate time series data with measurement error. The combination of latent factors with a vector autoregressive (VAR) model leads to the dynamic factor model (DFM), in which dynamic relations are derived within factor series, among factors and observed time series, or both. However, two limitations exist in the current DFM representatives and estimation: (1) the dynamic component of DFM contains either directed or undirected contemporaneous relations, but not both, and (2) selecting the optimal model in exploratory DFM is a challenge. Our paper serves to advance and evaluate DFM with hybrid VAR representations, which would then be estimated using LASSO regularization under the Structural Equation Model framework. This approach allows for the selection of the optimal hybrid dynamic relations in a data-driven manner. A simulation study is presented to investigate the sensitivity of finding the true hybrid dynamic relations in the structural model and the specificity of excluding the false relations using the LASSO-regularization versus the pseudo-ML approaches. We aim to offer guidance on model selection and estimation in person-centered dynamic assessments.

Keywords Time series data · Hybrid unified SEM · Regularized SEM · Dynamic factor model · Model-implied instrumental variable two-stage least square

A. Ye (✉)
Lehrstuhl für Psychologische Methodenlehre & Diagnostik, Department Psychologie, Ludwig-Maximilians-Universität München, München, Germany
e-mail: ai.ye@lmu.de

K. A. Bollen
Department of Psychology and Neuroscience, Department of Sociology, Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_28

1 Introduction

With the development of technology to collect intensive longitudinal or time series data (TSD), recent decades have been witnessing a surge of psychological and neurological research at the individual level. Studies that focused on person-specific dynamic assessment emphasize individual differences in individual characteristics and development. Within the psychometric field, researchers have developed psychometric modeling frameworks to fit traditional time series models, such as the Vector Autoregressive (VAR) models (Hamilton, 1994; Lütkepohl, 2005). There are growing number of psychometric literature on modeling individual dynamic models for TSD on a manifest level (e.g., Epskamp et al., 2018; Gates et al., 2019; Ye et al., 2021). For instance, two representative approaches are the unified Structural Equation Model (uSEM; Gates et al., 2010; Kim et al., 2007), as a time series extension of the SEM, and the graphical VAR (gVAR; Epskamp et al., 2018) model, as a time series extension of the network psychometric model. Recently, researchers have extended the VAR model into a hybrid representation (called 'hybrid VAR' by Molenaar & Lo. (2016), or 'hybrid uSEM' by Ye et al. (2021)) that can handle both the direct causal effects and undirected contemporaneous associations (Molenaar & Lo., 2016; Ye et al., 2021). These approaches differ in the specific variant of VAR representation and in the estimation framework to select and identify the optimal model. For instance, uSEM is usually identified by step-wise model search algorithms (Gates & Molenaar, 2012), while gVAR (Epskamp et al., 2018) or hybrid uSEM (Ye et al., 2021) adopt some variants of machine learning methods (e.g., regularization) to select and estimate the identifiable optimal sparse model.

However, so far these studies have focused on models with only manifest variables without accounting for measurement error. In practice, it is common that more than one, sometimes many, indicators measure the same underlying dynamic latent variable. When multiple indicators are available, latent time series variables could be formed to adjust for measurement error and to reduce the dimensions of the observed variables. The combination of the factor model and the time series model result in what we call the dynamic factor model (DFM; Browne & Nesselroade, 2005; Molenaar, 1985). In DFM, dynamic relations (e.g., lagged and contemporaneous relations) are allowed either within the factor series or amongst the factor and the observed time series. In fact, current dynamic modeling approaches include a factor model within their restricted VAR version to estimate a DFM. For example, the uSEM model in the GIMME framework (Gates & Molenaar, 2012) has been extended to the uSEM with latent variables (i.e., LV-uSEM), which they call the Latent Variable GIMME (or LV-GIMME; Gates et al., 2019). LV-GIMME estimates a DFM as a LV-uSEM using the stepwise model building algorithm, and an option to estimate parameters by use of the pseudo-maximum likelihood (i.e., pseudo-ML; Molenaar & Nesselroade, 1998).

However, the estimation of DFM with the more flexible VAR representation remains to be developed. Therefore, the primary purpose of our paper is to extend the regularized hybrid uSEM to the regularized hybrid uSEM with latent variables,

so that we can estimate a sparse DFM that allows for hybrid contemporaneous dynamic relations between the latent factors. Two steps address this overarching goal: the first is to reform the structural model of the latent variable uSEM (LV-uSEM) to its hybrid uSEM version, which the authors refer to as the latent variable hybrid uSEM (or LV-huSEM); the second is to perform model selection using the LASSO regularization in the search for the optimal sparse LV-huSEM; To evaluate the proposed method with existing ones, a simulation study will be conducted to compare 1) model recovery performance (sensitivity and specificity) of the LASSO regularization versus the pseudo-ML based stepwise model build when they are applied under the LV-huSEM context.

2 The Current Study

A general DFM for a single multivariate TSD is defined by two components, the measurement model and the structural or latent variable model (Molenaar, 1985). In the measurement model, LV-uSEM adopts a first-order processing factor series. This part adopts a confirmatory factor approach to obtain latent variables with the same qualitative meaning. For each individual, let $Y_t = [y_{1t}, y_{2t}, \ldots, y_{pt}]^T$ denote a vector of a p-variate time series at a given time point t, with $t = 1, \ldots, T$. Assuming Y_t represents a weakly stationary linear time series (i.e., with a constant mean, variance and covariance function). To ease the presentation, it is assumed that all the time series have zero mean function:

$$Y_t = \boldsymbol{\Lambda}\eta_t + \boldsymbol{\epsilon_t}, \boldsymbol{\epsilon_t} \sim N(\mathbf{0}, \boldsymbol{\Theta}). \tag{1}$$

In the structural model, LV-uSEM inherits a uSEM structure (Kim et al., 2007) that unifies temporal dependency and contemporaneous associations among latent factors:

$$\boldsymbol{\eta} = \boldsymbol{B}\eta + \boldsymbol{\zeta}, \boldsymbol{\zeta} \sim N(\mathbf{0}, \boldsymbol{\Psi}). \tag{2}$$

where $\boldsymbol{\eta} = [\boldsymbol{\eta_{t-1}}, \boldsymbol{\eta_t}]$ is a $2q \times T$ matrix. The variables are time-embedded by appending the data at $t - 1$ to that of t, thus expanded to two consecutive time points.

The LV-uSEM operates under the SEM framework for DFM that uses a pseudo-ML based model building algorithm and parameter estimation. Ideally, the specification for the structural model (1) should be guided by a priori theory. Unfortunately, very little is known about the individual dynamic pattern (Nichols et al., 2014; Wright et al., 2015, e.g.,). GIMME uses a data-driven forward selection algorithm where for every individual, it starts with a null model, and one path with the highest and significant modification indices is added iteratively until the model arrives at an acceptable fit (Gates et al., 2010). This model building procedure is in the open-source R package `gimme` (Lane et al., 2019).

However, there are two major ares that we propose to extend the individual modeling under the GIMME framework. First, it is imperative to move from the restrictive VAR representation in the DFM from a uSEM to the more flexible hybrid uSEM. That is, directed regressions and undirected error covariances among contemporaneous latent factor variables should be incorporated simultaneously. Because they can carry different causal interpretations as well as practical implications. Therefore, the first goal is to extend the structural model in LV-uSEM to the hybrid representation, i.e., LV-huSEM, by freeing up the parameters for the covariances of contemporaneous variables in the residual covariance matrix $\boldsymbol{\Psi}$ in Eq. 2.

Second, the forward selection method of model building is highly dependent on the starting model and the intermediate steps, and can arrive at an arbitrary final model. Results from the simulation study in Ye et al. (2021) also showed that this approach tends to miss relations with moderate to medium strengths even with the correct starting model and a large sample size. Regularization, in contrast, is a global, continuous model selection and a simultaneous estimation method. When using LASSO (Tibshirani, 1996), the sum of the absolute values are shrunken towards zero as λ increases, until they eventually reach exactly zero. Therefore, the current method seeks to replace the pseudo-ML stepwise searching and sequential estimation with the LASSO regularization for a simultaneous estimation of the extended LV-huSEM.

2.1 The Simulation Design

We designed a Monte Carlo simulation study to evaluate LASSO regularization and the pseudo-ML approach with respect to model recovery under the LV-huSEM context. The goal is to investigate the extent to which building LV-huSEM models with LASSO regularization is superior to the LV-GIMME approach in terms of (1) sensitivity of finding the true dynamic relations in the structural model, and (2) the specificity of excluding the false dynamic relations.

The Data Generating Model (DGM) The DGM is a five-factor DFAS with lag-1 effects and hybrid contemporaneous relations among the latent factors. In the measurement model, each factor has three strong to moderate indicators with no cross loadings or lagged relations. We include paths of different types and magnitudes to investigate the path recovery for hybrid dynamic relations. This includes a lag-1 autoregressive process within each factor, and a cross-lagged effect from a lagged factor to a contemporary factor. For the contemporaneous relations, we included direct paths as well as covariances. We varied the magnitude of coefficients and covariances to see whether these impact the recovery of the true DGM. We do not claim that the combination of these parameter values returns a common LV-huSEM model in practice.

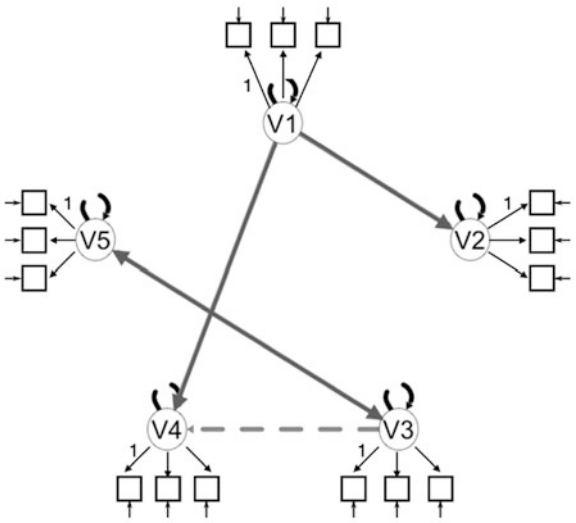


Fig. 1 DFM: a time-invariant five-factor DFAM with a hybrid uSEM structure

To investigate the influence of sample size on the performance, data is generated from the same DGM using time lengths varying from 60, 200, to 1000, representing a range from small to large in practice. All the DGMs will be replicated 1000 times, resulting in 3000 datasets. A weak stationary test will be performed on the data generating process, i.e., we will test that all eigenvalues of Φ have modulus less than one (Lütkepohl, 2005). All analyses will be performed in R, codes will be released and made publicly available on the Open Science Framework (OSF) (Fig. 1).

2.2 Analytic Procedure

For the pseudo-ML approach, confirmatory five-factor measurement models are estimated by pseudo-ML in *lavaan*, and factor scores are obtained by the default regression method of the 'lavPredict' function in *lavaan*. These factor score series will enter the subsequent structural model for model selection using pseudo-ML forward search in the GIMME package, function *indSEM*. The difference from the original setting in LV-GIMME is that here the starting structural model is a huSEM (with the covariance matrix $\boldsymbol{\psi}^*$) instead of the more restricted uSEM (with $\boldsymbol{\psi}$). Additionally, we focus on individual models only, i.e., no group level model is involved. For this reason, we refer to this method "pseudoML-FS-huSEM" to indicate that it uses modification indices for the search of sparse huSEM models using the factor scores.

For the proposed method, the LV-huSEM under LASSO regularization (i.e., LASSO-LV-huSEM) will be implemented under the regularized SEM framework. After the LV-huSEM model structure is specified in *lavvan*, *regsem* can import

the *lavvan* output and perform LASSO regularization with the user-defined list of parameters in the penalty function $\lambda P(\theta^*)$. To ensure that factor series represent latent constructs that are consistent with those of "pseudoML-FS-huSEM", the same confirmatory factor structure is estimated without penalty. Factor loading parameters belong to the freely estimated set in θ but not in the regularization set θ^*. Parameters in set θ^* are regression coefficients for cross-lagged effects and contemporaneous effects except the AR coefficients, as well as the error covariance among contemporaneous latent factors. Ideally, the optimal λ (with the lowest BIC) penalizes all unnecessary parameters to zero and estimates the remaining ones, unraveling the true type of relation between any two latent factors from five possibilities: two cross-lagged effects, two contemporaneous regression coefficients, and one contemporaneous error covariance.

We use sensitivity and specificity to evaluate the accuracy of recovering relations with the correct direction. Sensitivity represents the power to detect true relationships; specificity, in comparison, represents the percentage of non-existing paths in the DGM that the search procedure accurately omitted in the final model. These measures allow for an evaluation of a model's ability to detect true recovery and to reject false ones.

3 Result

Let us first turn to the sensitivity for recovering true relations from the starting LV-huSEM (Fig. 2). Both methods showed an excellent sensitivity for lag-1 effects regardless of the sample size. Besides lag-1 relations, the probability to recover another true path by any method depends largely on the sample size: the recovery rates were low when the sample size was small (N = 60), overall acceptable at a medium sample size (N = 200) and satisfactory given a large sample size (N = 1000). Specifically, between the two methods, LASSO-LV-huSEM showed an overall higher sensitivity to strong relations (i.e., directed, covariance, and cross-lagged relations) when given a medium or large sample size. Surprisingly, pseudoML-FS-huSEM performed poorly in recovering the strong directed path even with a large sample size. A closer examination revealed that the majority of time the model tended to recover a true strong directed path as a covariance relation and sometimes as a reversed sign directed path (hence a high rate of direction false positive). This is a scenario of a recovery that counted as a "path presence recovery" but not as a "direction recovery" in Ye et al. (2021).

We also examined the path-specific "direction false positive rate", defined as the proportion of replication in which a true association was recovered with the wrong direction. Not surprisingly, it was observed that some relations were recovered with a wrong direction when the sample size was small. However, even when the sample size was sufficient and the true path was recovered, sometimes additional paths might still be selected when there existed a strong correlation between the two variables. Hence, direction false positive rates did not necessary go

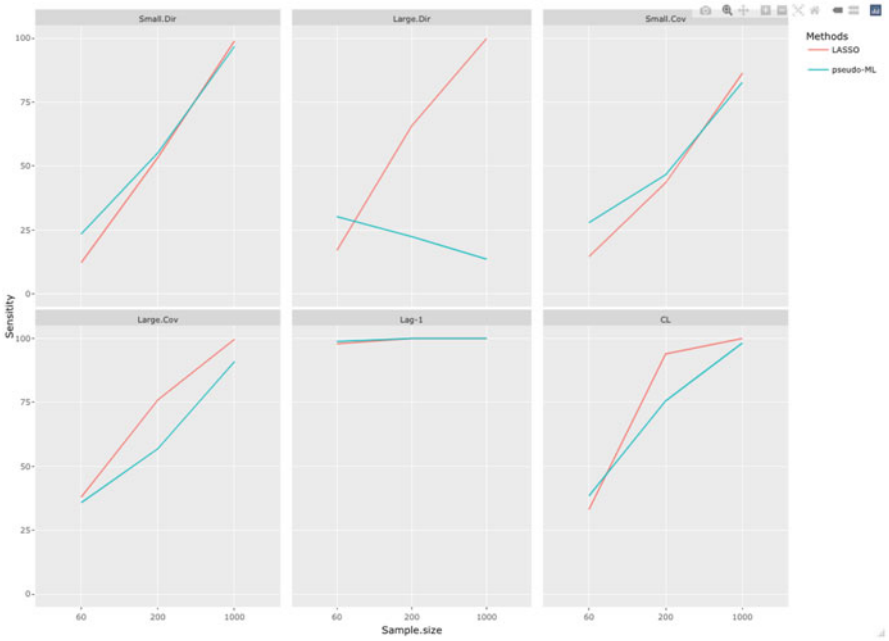


Fig. 2 Sensitivity of path recovery by path type and strength across sample size. Note: Small.Dir = small directed path, Large.Dir = large directed path, Small.Cov = small covariance relation, Large.Cov = large covariance relation, CL = cross-lag effect, Lag-1 = lag-1 effect, Large.FL refers to factor loadings of 0.9, Med.FL refers to factor loadings of 0.7

down with the increase of sample size. Overall, except for cross-lagged relations, pseudo-ML methods had higher direction false positive rates in relation to the true paths in the DGM than the LASSO methods. This is partly the reason that pseudoML-FS-huSEM had very poor sensitivity under some conditions. That is, some relations were recovered only with a wrong direction or type of relation. More problematically, pseudo-ML showed a high chance (67% at N = 200 or 96% at N = 1000) of recovering a reversed signed directed path or a covariance when there exists a strong directed path. These consistent observations (that all methods using the factor scores showed a higher rate of direction false positive than their counterparts) suggested that the issue of a wrong direction recovery is very likely tied to the use of factor scores in place of the latent variables.

Finally, both methods reached a path specificity above 90%, suggesting they are reliable in rejecting false paths that were unrelated with those pairs of variables that have a true relation of another form or direction. However, the direction specificity (i.e., the odds of ruling out any path when it is truly false) dropped substantially for pseudoML-FS-huSEM (around 72–77%) or any method that used factor scores. This is again direction false positive paths.

4 Discussion

The current study serves to advance the model search and estimation for a DFM with a hybrid VAR representation. Two goals were achieved in the proposed framework. First, we extended the structural model of the latent variable uSEM (LV-uSEM) to its hybrid uSEM version, i.e., the latent variable hybrid uSEM (or LV-huSEM). In this way, the extended LV-huSEM estimates a dynamic factor model with a hybrid VAR representation in the structural model. Second, LASSO regularization is used to perform both model selection and estimation for the optimal sparse latent variable hybrid uSEM. Compared to the current approaches, where the measurement model and the structural model are estimated sequentially with a stepwise model search procedure using factor scores obtained prior to the model selection (e.g., LV-GIMME; Gates et al., 2019), the current method provides a model search on a continuum and a simultaneous estimation without calculating factor scores. A simulation study was conducted to investigate to what extent the novel estimation method for the LV-huSEM models is superior to the pseudo-ML approach similar to the individual model in the LV-GIMME framework with respect to model recovery performance.

For model recovery, we found that the two approaches have comparable recovery rates for some relations such as lagged effects and moderate contemporaneous effects among factors, and they both are reliable in recovering a close-to-true structural model when the sample size is medium to large. However, the pseudo-ML methods using factor scores have a higher chance to commit a direction false positive on strong directed relations, that is, a tendency to recover a strong directed path as one with a reversed direction or as an undirected covariance relation. The result suggests that the use of factor scores instead of latent variables is more likely to select a model with a higher false positive rate.

The simultaneous analysis using LASSO regularization under the LV-huSEM is easy to implement and can avoid biases from the use of factor scores, but it might be more limited in the size and complexity of the model than sequential analyse like the GIMME approach. This is because the use of factor scores reduces the dimension of the parameter space. Optimizing the covariance matrix of observed variables with a higher dimension is more difficult than that of the latent factors. Our DGM might represent an over-simplified, over-sparse DFM, with a very standard measurement structure without cross-loadings or local dependency structures. In practice, the latent variable relations in a DFM could be much denser with many weak to moderate relations. Another aspect that is out of the scope of the current study is DFMs for multiple subject time series. The focus of this study is on individual dynamic models, for which there is no consideration of between-person effects nor attempt to aggregate individual models. However, the use of group-level or between-person information (i.e., similarities and variances across individuals) has been shown as an effective way to extract true effects from noise so that it avoids the risk of over-fitting individual dynamic models (Asparouhov et al., 2018; Gates et al., 2019).

Our results of the LASSO regularized hybrid uSEM with latent variables highlights the flexibility of the LASSO regularized SEM in estimating individual DFMs. The data-driven LASSO penalty opens up a variety of possibilities in the development and appraisal of individual dynamic theories. The penalization structure relies on which part of the model is more supported by theory, and which part is more uncertain and needs to be explored by the data. The flexibility of regularization with user-defined estimation and penalization structure lifts the dichotomous boundary between the exploratory approach and the confirmatory one and allows for an expansion and refining of theory on a continuum.

References

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(3), 359–388. Retrieved from https://doi.org/10.1080/10705511.2017.1406803

Browne, M. W., & Nesselroade, J. R. (2005). Representing psychological processes with dynamic factor models: Some promising uses and extensions of autoregressive moving average time series models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 415–452). Lawrence Erlbaum Associates Publishers.

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018). The gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research, 53*(4), 453–480.

Gates, K. M., Fisher, Z. F., & Bollen, K. A. (2019, 06). Latent variable GIMME using model implied instrumental variables (MIIVs). *Psychological Methods, 25*, 227.

Gates, K. M., & Molenaar, P. C. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage, 63*(1), 310–319.

Gates, K. M., Molenaar, P. C., Hillary, F. G., Ram, N., & Rovine, M. J. (2010). Automatic search for FMRI connectivity mapping: an alternative to Granger causality testing using formal equivalences among SEM path modeling, VAR, and unified SEM. *NeuroImage, 50*(3), 1118–1125.

Hamilton, J. D. (1994). *Time series analysis* (Vol. 2). Princeton, NJ: Princeton University Press.

Kim, J., Zhu, W., Chang, L., Bentler, P. M., & Ernst, T. (2007). Unified structural equation modeling approach for the analysis of multisubject, multivariate functional MRI data. *Human Brain Mapping, 28*(2), 85–93.

Lane, S., Gates, K. M., Fisher, Z. F., Arizmendi, C., Molenaar, P. C., Hallquist, M., et al. (2019). Gimme: Group iterative multiple model estimation [Computer software manual]. Retrieved from https://github.com/GatesLab/gimme/ (R package version 0.6-1)

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

Molenaar, P. C. (1985, June). A dynamic factor model for the analysis of multivariate time series. *Psychometrika, 50*(2), 181–202. Retrieved from https://ideas.repec.org/a/spr/psycho/v50y1985i2p181-202.html

Molenaar, P. C., & Lo, L. L. (2016). Alternative forms of Granger causality, heterogeneity and non-stationarity. In W. Wiedermann & A. von Eye (Eds.), *Statistics and causality: Methods for applied empirical research* (pp. 205–230). Wiley.

Molenaar, P. C., & Nesselroade, J. R. (1998). A comparison of pseudo-maximum likelihood and asymptotically distribution-free dynamic factor analysis parameter estimation in fitting covariance-structure models to block-Toeplitz matrices representing single-subject multivariate time-series. *Multivariate Behavioral Research, 33*(3), 313–342. Retrieved from https://doi.org/10.1207/s15327906mbr33031 (PMID: 26782717)

Nichols, T. T., Gates, K. M., Molenaar, P. C., & Wilson, S. J. (2014). Greater bold activity but more efficient connectivity is associated with better cognitive performance within a sample of nicotine-deprived smokers. *Addiction Biology, 19*(5), 931–940.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288.

Wright, A. G., Beltz, A. M., Gates, K. M., Molenaar, P. C., & Simms, L. J. (2015). Examining the dynamic structure of daily internalizing and externalizing behavior at multiple levels of analysis. *Frontiers in Psychology, 6*, 1914.

Ye, A., Gates, K. M., Henry, T. R., & Luo, L. (2021, June 01). Path and directionality discovery in individual dynamic models: A regularized unified structural equation modeling approach for hybrid vector autoregression. *Psychometrika, 86*(2), 404–441. Retrieved from https://doi.org/10.1007/s11336-021-09753-6

Optimizing Multistage Adaptive Testing Designs for Large-Scale Survey Assessments

Usama S. Ali, Peter W. van Rijn, and Frederic Robin

Abstract Multistage adaptive testing (MST) is used by several large-scale survey assessments to enhance measurement efficiency and improve test-taker experience. In large-scale survey assessments, the target populations (e.g., countries in an international comparative assessment) can be of different proficiency levels. Accordingly, when designing an MST, multiple competing goals need to be considered: (a) to better match the proficiency level of a given respondent and the difficulty of assessment; and (b) to get sufficient quality data (i.e., enough responses per item across the proficiency levels) to support the estimation of item parameters. In developing the instruments for large-scale survey assessments, it is critical to ensure content coverage as well as use of the full item pool. Seeking an optimal design for the Programme for the International Assessment of Adult Competencies (PIAAC), in this paper we address comparing the performance of different designs in terms of various evaluation criteria and demonstrate instrument development using optimization methods. We also discuss results and implications.

Keywords Stepwise assembly · Large-scale survey assessments · Multistage adaptive testing · Optimal methods · Programme for the International Assessment of Adult Competencies

U. S. Ali (✉)
Educational Testing Service, Princeton, NJ, USA

South Valley University, Qena, Egypt
e-mail: uali@ets.org; usama.ali@edu.svu.edu.eg

P. W. van Rijn
ETS Global, Amsterdam, The Netherlands
e-mail: pvanrijn@etsglobal.org

F. Robin
Educational Testing Service, Princeton, NJ, USA
e-mail: frobin@ets.org

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_29

1 Introduction

Digital delivery has become the default assessment mode in this era in national and international large-scale survey assessments (LSAs). Such digital delivery enables competencies to be measured more efficiently and effectively, allows the measurement of new competencies, and makes the collection and analysis of process data possible. Digital delivery also permits for more personalized assessments via adaptive tests (Bennett, 2018). Such adaptive testing leads to a better match between test difficulty and student proficiency and enhances measurement precision across the entire proficiency distribution when compared to linear tests. Multistage adaptive tests (MSTs) have became so popular that several LSAs have adopted or are considering adopting MSTs in order to gain advantages such as improved test taker experience. For instance, the Programme for International Assessment of Adult Competencies (PIAAC) has used an MST design since 2012 and the Programme for International Student Assessment (PISA) since 2018 (Yamamoto et al., 2018). In LSAs, the target populations (e.g., countries in an international comparative assessment or regions in a national assessment) can be of different proficiency levels. PIAAC, as one of the largest and most innovative international assessments focusing on measuring knowledge, skills, and attributes in adult populations, is implemented across 38 countries in more than 50 languages (Kirsch & Lennon, 2017).

In designing an MST system, multiple competing objectives need to be considered. Among these is to match the proficiency level of a given respondent and the difficulty level of the assessment better than can be accomplished with a linear testing system. Another important objective is to get sufficient quality data (i.e., enough responses per item collected from respondents of different proficiencies across the proficiency continuum) to support the estimation of item parameters and to properly detect any item misfit at the national level. Furthermore, in developing the testlets for an MST system, it is critical to ensure content coverage and use the full item pool in addition to other constraints (e.g., considering the assessment time limit). Hence, the assembly of the assessment instruments in this context is an optimization problem that is subject to a wide range of constraints and research is required to identify designs best suited for LSAs.

The paper is outlined as follows. In the next section, we illustrate a stepwise assembly approach along with the various constraints to be considered. We then demonstrate this assembly framework via an application of the proposed MST design of PIAAC. We compare the performance of different designs in terms of various evaluation criteria.

2 Stepwise Assembly

Our proposed stepwise assembly approach splits the optimization problem outlined above into multiple manageable steps where each step has its own objectives and constraints (van der Linden, 2005). Let us assume an MST design with m stages. In our assembly approach, the item pool is to be partitioned into m mutually exclusive sets of items. Each set is then used to assemble testlets (i.e., modules or item blocks) for each stage (e.g., medium difficulty testlets for the routing stage and of varying levels of difficulty for the subsequent adaptive stages). This approach lends itself to build a balanced MST design that rotates all items across all relevant test positions (i.e., stages). Also, this approach allows for developing linear tests, where only the testlets assembled for the routing stage from each item set are delivered across all testing stages. Furthermore, hybrid designs can be specified so that some proportion of test takers is randomly assigned to the linear tests and the others are assigned to the MST (e.g., this is done in PISA 2022 mathematics and PIAAC cycle 2 literacy and numeracy designs). Linear administration to some subsample can help to facilitate item calibration by ensuring sufficient item-level data and prevents the need for misrouting.

In the assembly practices for PIAAC and PISA, we consider the constraints related to testlet length, linking (trend) versus new items, content specifications (e.g., context, process), item format (e.g., simple multiple-choice, complex multiple-choice), scoring (e.g., human versus machine scored), the unit structure (i.e., items from the same unit are always administered together), and timing information. In addition, constraints on the amount of overlap between testlets and item pairs can be added to enhance the efficiency of the item calibration. The following steps generally follow the optimization methods described in van der Linden (2005)

The stepwise assembly approach can be summarized in the following main steps, where the binary decision variables are different for each step:

(a) Item set assembly:
$x_{jt} = 1$ if unit j is in item set t.
(b) Testlet assembly:
$x_{jt} = 1$ if unit j is in testlet t.
(c) Multistage path assembly:
$x_{jt} = 1$ if testlet j is in path t.
(d) Linear form assembly:
$x_{jt} = 1$ if testlet j is in form t.

The objective function and constraints are different for each step as well. Regarding the target function in the testlet assembly step, both information and expected score functions can be used. Under conditional independence, the unit information function $I_j(\theta_i)$ is the sum of item information functions in unit j evaluated at θ_i. A target test information function $\mathcal{I}(\theta_i)$ can be specified. Then, the minimax principle can be used to minimize $\epsilon \geq 0$ subject to

$$\mathcal{I}(\theta_i) - \epsilon \leq \sum_{j=1}^{J} I_j(\theta_i) x_{jt} \leq \mathcal{I}(\theta_i) + \epsilon, \quad \text{for all } i \text{ and } t.$$

Similarly, the unit characteristic curve (i.e., the expected score as a function of θ) is the sum of item characteristic curves. A target test characteristic function $\mathcal{T}(\theta_i)$ can be used in a similar fashion (Ali & van Rijn, 2016).

For the testlet assembly step, other binary decision variables can be added to control item pairs and overlap (connectedness; Eccleston & Hedayat, 1974). For controlling item pairs, we can let $y_{jj't} = 1$ if units j and j' are in testlet t with $j < j'$. For item overlap, we can introduce $z_{jtt'} = 1$ if unit j is in testlet t and t' with $t < t'$. However, when using pairs of units and pairs of testlets, the number of decision variables can quickly become very large.

For each testlet t, constraints of category c can be formulated as:

$$n_c^{\min} \leq \sum_{j=1}^{J} n_{cj} x_{jt} \leq n_c^{\max}, \quad \text{for all } c \text{ and } t.$$

Constraints of this type can be categorical ones such as item format or quantitative ones such as expected response time. Other constraints (e.g., "enemy" units) can be added as well.

3 Application

We set up a simulation study based on the PIAAC cycle 2 main study design proposed for literacy and numeracy measures, see Fig. 1. It is a hybrid multistage adaptive/linear design. To address the uncertainty in item parameters, 25% of respondents are assigned to linear (non-adaptive) tests consisting only of testlets prepared for the routing stage where routing is therefore not dependent on the performance of test takers during the assessment. Stage 1 is the locator where all respondents receive the same set of items (e.g., eight numeracy items). Regardless of their assignment to the adaptive or linear routes, all respondents are randomly assigned to Design A or Design B to balance the item positions between Stages 2 and 3. This design depicted in Fig. 1 allowed us to study three different designs by changing the percentage of respondents that take the linear tests: the linear design (100%), the MST design (0%), and the hybrid design (25%).

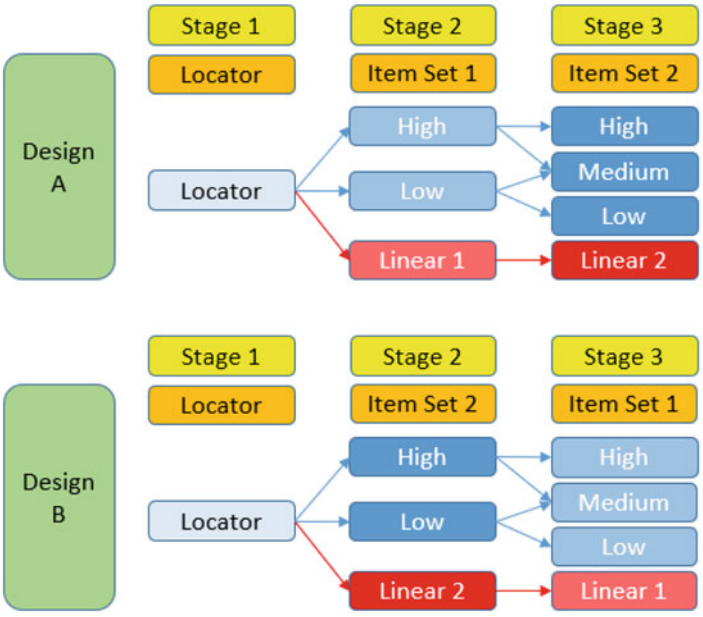


Fig. 1 PIAAC literacy and numeracy hybrid MST design

3.1 Simulation Study

In the simulation study, we took the following steps:

1. We used the PIAAC numeracy item pool consisting of 72 items (structured in 44 units) and their estimated item parameters (see Table 1) to develop Stages 2 and 3 testlets for the different designs. In addition to these items, the Stage 1 testlet with eight locator items is designed to be the same across all studied designs.
2. We selected seven groups out of 28 countries to represent the heterogeneous PIAAC populations including both the least able group (mean = 206, standard deviation [SD] = 59) and the most able one (mean = 288, SD = 44) based on cycle 1 published means and standard deviations, see Table 2 for the selected groups. Then, the ability parameters were generated for each simulee in each of these simulated groups. Each group consisted of 3000 simulees. This sample size was chosen to reflect the sampling design and is actually the expected target sample size.
3. Given that we were only assembling the testlets for Stages 2 and 3, we partitioned the 72-item pool into two item sets, each with 36 items (22 units), then assembled the following testlets (12 items per testlet) using mixed-integer linear programming (Diao and van der Linden, 2011):

Table 1 Item pool statistics (72 items)

Parameter	Minimum	Mean	SD	Maximum
a	0.42	0.97	0.29	1.65
b	−1.62	0.32	0.94	3.62

Table 2 Mean and standard deviation of the seven studied groups

Group	1	2	3	4	5	6	7
Mean	206	247	256	265	273	278	288
SD	59	50	54	56	46	51	44

(a) six adaptive Stage 2 testlets (targeting two levels of difficulty: three low-difficulty and three high-difficulty testlets) from each item set. The low- and high-difficultly testlets are centered around the middle of numeracy proficiency Level 2 and middle of proficiency Level 3 cut scores[1] on the PIAAC scale, respectively.
(b) six adaptive Stage 3 testlets (targeting three levels of difficulty: two low-difficulty, two medium-difficulty, and two high-difficulty testlets) from each item set. The low-, medium-, and high-difficultly testlets are centered around the middle of proficiency Level 2, Level 3, and middle of Level 3 cut scores on the PIAAC scale, respectively.
(c) six linear testlets of medium difficulty from each item set.

4. Based on the known item parameters and simulees ability parameters, item responses were generated using the two-parameter logistic model (e.g., Lord, 1980). The datasets were created for three different designs: linear, MST, and hybrid (where 25% of the simulees take the linear test).

To evaluate the results across the different designs, we used a set of evaluation criteria. We checked item exposure (i.e., the number of responses per item) and the ratio of the maximum and minimum item exposure for each design. Since expected a posteriori (EAP) estimates of the proficiency (θ) and their variances (VAP; variance a posteriori) are used to estimate latent regression parameters in group-score assessments (see e.g., Thomas, 1993), we use them to evaluate the measurement precision in the different designs. Item response theory (IRT)-based reliability was calculated using Kim's (2012) formula:

$$R = \frac{\text{Var(EAP)}}{\text{Var(EAP)} + \text{Mean(VAP)}} . \tag{1}$$

Additionally, the relative efficiency (RE) of the EAP estimates can be used to compare the designs with respect to measurement precision and item calibration. The relative efficiency with respect to the EAP can be determined using the ratio of their variances (VAP) following the approach outlined by Yamamoto et al. (2019),

[1] For more details about description of PIAAC proficiency levels in each domain, refer to the technical report (OECD, 2019).

which we refer to as *relative EAP-efficiency*:

$$RE_{\text{EAP}}(\theta) = \frac{\text{VAP}_2(\theta)}{\text{VAP}_1(\theta)} \,. \tag{2}$$

3.2 Selection of Findings

Figure 2 shows the distribution of the simulees across the four adaptive paths within the MST design for each of the seven groups. The distribution changes with increasing mean PIAAC numeracy scores as more simulees are being routed to paths of higher difficulty.

Figure 3 shows boxplots for the number of observed responses per item in each of the seven groups for each of the three studied designs. It is clear that the linear design leads to a uniform distribution of item responses per item across groups. The expected number of responses per item in the linear design is 1000 (given that Design A and Design B each had 1500 respondents, with each respondent taking 24 out of 72 items). The expected minimum number of responses per item in the hybrid MST design is 250 (given that one out of four simulees was randomly assigned to the linear test). Such data (e.g., minimum sample size per item) in the hybrid design

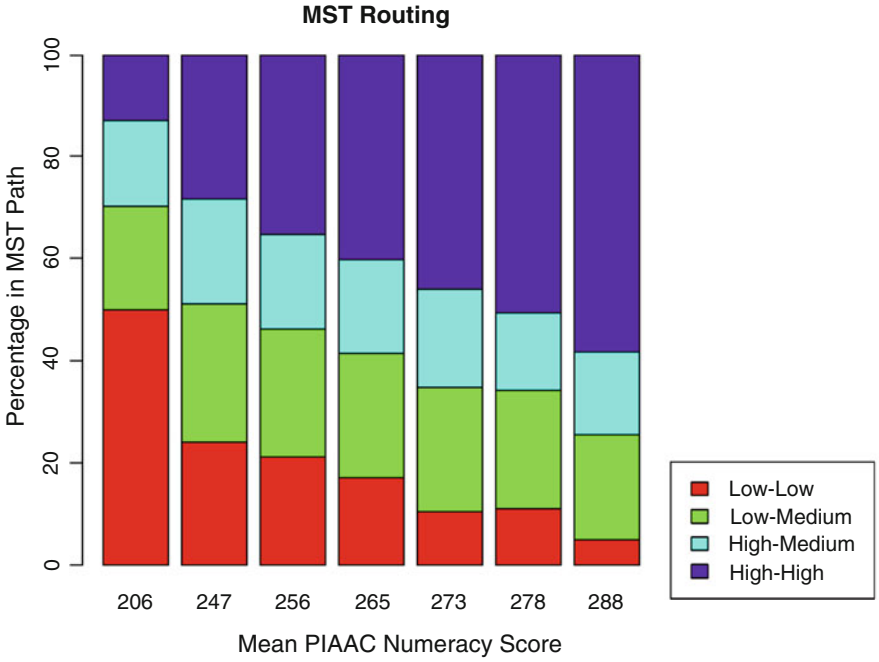


Fig. 2 MST routing across the studied groups

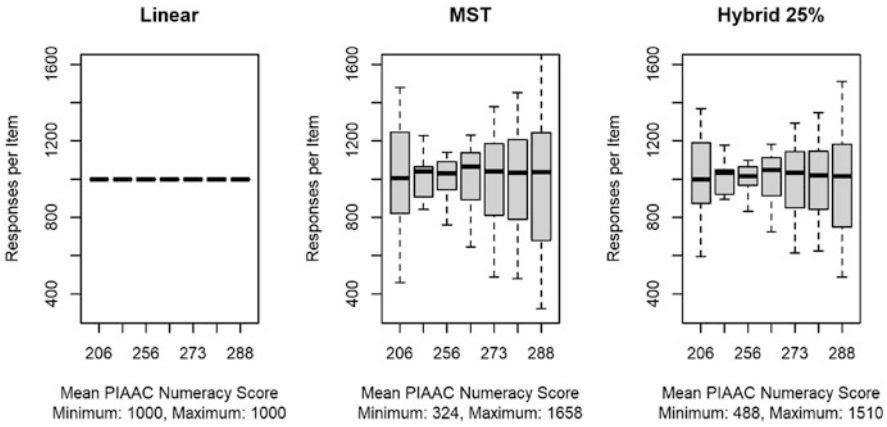


Fig. 3 Distribution of number of responses per item across groups by design

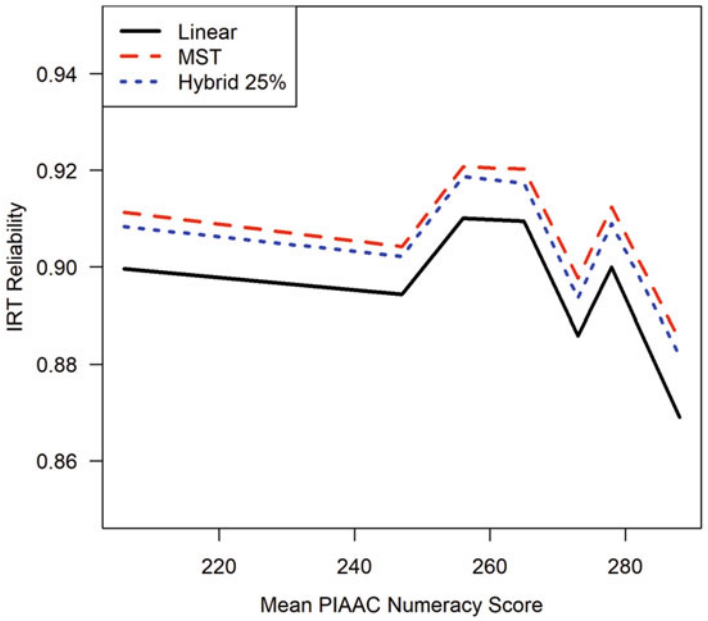


Fig. 4 Average IRT reliability of EAP estimates for the studied groups by design

would be sufficient to detect potential item-by-country interactions and recover item parameters. As illustrated by Fig. 3, the ratios of maximum to minimum exposure are: 1 for the linear design, 5.1 for the MST design, and 3.1 for the hybrid design.

Figure 4 displays the IRT reliability as a function of the mean numeracy score for each of the designs. The curves are as expected, with the MST design providing

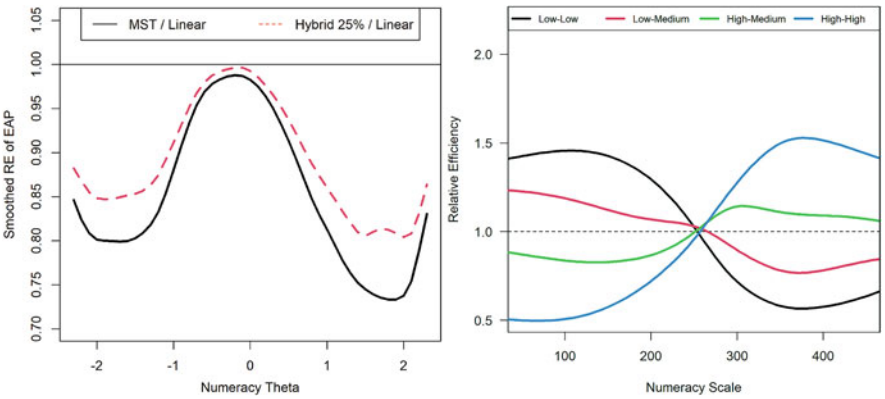


Fig. 5 Smoothed relative EAP-efficiency (*left*) and relative efficiency (*right*) of MST and hybrid MST designs to the linear design (Note that the horizontal line at 1.0 in each panel refers to the threshold of equal efficiency.)

the highest reliabilities with some variation across groups due to the differences in their variances.

The left panel of Fig. 5 shows the smoothed relative efficiency of the EAP estimates comparing the MST and hybrid designs with the linear design. For extreme proficiencies, average precision gains for the MST design is up to 37% with slightly less gain for the hybrid design. The right panel provides the relative efficiency as defined by the ratio of the average test information function of the linear design divided by the average test information function for the different routing paths of the MST design. The maximum gain in measurement precision can reach up to about 50%.

4 Discussion

The rapid increase in technology has allowed LSA to move towards more personalized tests that would improve the test-taking experience through matching person proficiency and test difficulty. The adaptive testing designs are also expected to improve measurement precision. In this paper, we presented a stepwise assembly approach used recently for developing the cognitive instruments for LSAs using automated test assembly methods. We shared the main steps for this approach to help address this highly constrained optimization problem. In summary, the assembly framework works well in developing the instruments to achieve the various competing goals necessary for LSAs. Furthermore, with the automated procedure we assign items to testlets satisfying several important constraints in a more efficient way than the manual assembly that was heavily dependent on content experts.

In the context of the PISA 2018 reading assessment, Yamamoto et al. (2019) found that measurement precision could be improved by 4–5% overall and up to 10% for extreme proficiencies with MST. This result supports the use of MST in LSAs. However, some limitations can be mentioned. For one, we did not include comparisons with other approaches for test assembly (e.g, on-the-fly assembly; Zheng & Chang, 2015). In addition, it would be of interest to study the impact of different designs on the final reporting results (e.g., plausible values, group statistics).

Regarding the recommended assessment design, the hybrid MST design does seem to provide a robust design to be implemented operationally by ensuring the construct representation in the assembled testlets, collecting sufficient data to recover item parameters and detect any potential item misfit, and increasing measurement precision. It also provides other features such as balancing item exposure and assessment time across respondents of varied proficiency levels. For future developments, other aspects to consider based on real data are checking levels of missing data (omit/not reached) that reflect the test-taking experience and level of engagement, and the quality of item parameter estimation. Other adaptive designs can be considered as well such as targeted testing (e.g., Mislevy & Wu, 1996) and targeted MST (e.g., Berger et al., 2019). Given that the context is group-score assessments, it is also critical to investigate how to optimize the design in order to get more accurate group-level inferences.

References

Ali, U. S., & van Rijn, P. W. (2016). An evaluation of different statistical targets for assembling parallel forms in item response theory. *Applied Psychological Measurement, 40*(3), 163–179.

Bennett, R. E. (2018). Educational assessment: What to watch in a rapidly changing world. *Educational Measurement: Issues and Practice, 37*(4), 7–15.

Berger, S., Verschoor, A. J., Eggen, T. J., & Moser, U. (2019). Efficiency of targeted multistage calibration designs under practical constraints: A simulation study. *Journal of Educational Measurement, 56*(1), 121–146.

Diao, Q., & van der Linden, W. J. (2011). Automated test assembly using lp_solve version 5.5 in R. *Applied Psychological Measurement, 35*(5), 398–409.

Eccleston, J., & Hedayat, A. (1974). On the theory of connected designs: characterization and optimality. *The Annals of Statistics, 2*, 1238–1255.

Kim, S. (2012). A note on the reliability coefficients for item response model-based ability estimates. *Psychometrika, 77*(1), 153–162.

Kirsch, I., & Lennon, M. L. (2017). PIAAC: A new design for a new era. *Large-scale Assessments in Education, 5*, 11.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing*. ETS Research Report RR-96-30, Educational Testing Service, Princeton, NJ.

OECD. (2019). *Technical report of the Survey of Adult Skills (PIAAC)* (3rd ed.). OECD Publishing.

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics, 2*(3), 309–322.

van der Linden, W. J. (2005). *Linear models for optimal test design*. Springer.

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement: Issues and Practice, 37*(4), 16–27.

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2019). Introduction of multistage adaptive testing design in PISA 2018. In *OECD Education Working Papers No. 209*. OECD Publishing.

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

Psychometric Modeling of Handwriting as a Nonverbal Assessment Instrument and Its Properties

Yury Chernov

Abstract In psychological practice, there are often conditions, when experts cannot use traditional psychometric instruments. That relates in particular to forensic and criminal psychology. In these cases, experts apply nonverbal instruments, to which belongs handwriting psychological analysis. Its major advantages: it is based on the normal person's activity and practically cannot be manipulated. However, the validity of nonverbal methods and procedures is typically poor. To improve that a method should be formalized. In the current work, we present formalized models of handwriting. The models are based on statistical regressions and are implemented in the HSDetect framework. Although handwriting features still must be evaluated manually, since the proper automated solutions are not available, the formalized character of the procedure ensures objectivity and transparency. The modeled traits, to which belong psychological characteristics, cognitive states, typological dimensions, or disease markers, are evaluated algorithmically. HSDetect includes several models, which psychometric properties allow the selection of a proper one in a particular case. The procedure has been successfully validated against several well-known psychometric tests. HSDetect was applied in several practical cases. In the current work, we present two: the screening instrument for of Alzheimer's disease and the model of aggressiveness.

Keywords Formalization · Handwriting analysis · Handwriting modeling · Psychometric properties · Forensic and criminal psychology · Alzheimer's disease · Aggressiveness

1 Introduction

In psychological practice, there are often conditions, under which experts cannot use traditional psychometric instruments. That relates in particular to forensic

Y. Chernov (✉)
Institute for Handwriting Sciences, Zurich, Switzerland
e-mail: yc@ihs-sgg.ch

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_30

and criminal psychology. The reasons for this are rather objective. For instance, the person under the expertise can be not available, or his/her answers should not be trusted due to the obvious tendency to manipulate them. In these cases, experts base their conclusions on nonverbal instruments. However, the validity of these instruments is typically poor. That is due to the subjective character of the methods. To make a method more objective and to improve its validity, it should be formalized: a formalized method allows for proper statistical research and validation studies. One such method is handwriting analysis.

Handwriting analysis has some advantages over questionnaire-based instruments. The major ones are that it is based on the normal person's activity and practically cannot be manipulated. In the current work, we present formalized models of handwriting and evaluate their psychometric properties. The models are based on statistical regressions and are implemented in the HSDetect computer-aided framework. Due to the formalized character, the application is highly objective and transparent. The procedure has been successfully validated both against several well-known psychometric tests and against expert evaluations.

Different models based on HSDetect were applied in several practical cases. In the current work, we present two examples. The first one relates to the markers of Alzheimer's disease in handwriting. The second one presents psychological construct aggressiveness.

Formalized handwriting analysis has three important properties that make it especially effective in certain cases. First, it provides wide coverage of personal characteristics. That makes it very flexible in the modeling of different psychological constructs. Secondly, it ensures the exclusion of social desirability. It could be thus a good supplement to the standard psychological assessment. Thirdly, it enables the creation and maintenance of a normative database. That allows not just abstract formal quantification, but modeling based on normative data.

2 Method

2.1 A Subsection Sample

The fact that handwriting might reflect a person's physiological and health status (Caligiuri & Mohammed, 2012; Harralson & Miller, 2018), as well as his/her permanent psychological characteristics (Seibt, 1994; Michel, 1982), is well known. The last was traditionally the area of graphology. However, traditional graphology lacks profound scientific background and sufficient validation. It has been often rightly criticized for this (Chamorro-Premuzic & Furnham, 2014; Kanning, 2019). Although graphology did have a long path of successful experiences, mainly in human resource assessment, it does not satisfy the actual requirements for a valid psychometric method. Graphology lacks systematization and objectivity. The work of an expert is not transparent. The less-structured and intuition-based procedure

makes the result very dependent on the expert. Besides, the typical outcome is a plain text, which depends mostly upon the ability of the expert to compile such texts, which become the subject of an ambiguous interpretation. However, they aim the criticism rather at the way graphology is being practiced, than at the idea of the relation between handwriting and traits itself.

A formalized handwriting analysis allows for solving the mentioned problems at the same time while keeping the positive historical experience, where it is appropriate. In the current article, we refer to the HSDetect framework (Chernov, 2011, 2021). It implements an analytical approach to handwriting modeling. HSDetect demonstrates promising validation results (Chernov, 2018; Chernov & Caspers 2020) against a series of psychometric tests.

The model includes three major objects: first, handwriting characteristics as independent variables. They are derived during the analysis and evaluation of a handwriting specimen. Second, traits as dependent variables and, third, so-called graphometric functions. The graphometric functions represent the relations between handwriting characteristics and traits. Under a trait, we understand any feature of a person, which can be reflected in his/her handwriting. That can be a psychological feature, e.g., aggressiveness, a dimension of a psychological typology or model, e.g., Big Five neuroticism, or a disease marker, e.g., Alzheimer's.

The evaluation of a handwriting specimen in HSDetect must be done manually. The available computer programs for the automatic evaluation of scanned texts (e.g. MovaAlyzeR,[1] CEDAR-FOX,[2] or Masquerade[3]) do not provide a proper solution. First, they can cover only very few simplest handwriting signs. Secondly, their reliability is very low. That relates both to the analytical approach and to the neural network solutions, proposed in many publications.

To adequately model handwriting, HSDetect includes about 200 handwriting signs and over 700 handwriting characteristics. Under a handwriting sign, we understand a general quality of handwriting, for example, letter size. A handwriting characteristic is a particular manifestation of a handwriting sign in the analyzed sample. Letter size can have the following five characteristics: medium (2–3 mm), small (1–2 mm), very small (<1 mm), large (3–5 mm), and very large (>5 mm).

To provide an objective, unambiguous and reliable evaluation of handwriting signs it is very important to formally define them. Therefore, for instance, for the size experts consider letters of the middle zone (a, c, e, m, n, o, u, v, w). Moreover, only inner letters in words, i.e. without the first and the last ones, are taken. Such definitions for every sign and characteristic are necessary. First, often handwriting experts do not agree on the evaluation of a particular specimen. Especially when it is complicated and signs are not consistent. Secondly, unambiguity and modeling require a quantitative evaluation. The formalized definition includes an algorithm of evaluation, which should work in all cases.

[1] www.neuroscript.net

[2] www.cedartech.com

[3] www.nitesrl.com

In HSDetect, both handwriting characteristics and traits are presented on a continuous scale from 0 to 1. A graphometric function is a linear regression. The trait level y is defined as a function of levels of handwriting characteristics xi:

$$y = \sum a_i \cdot x_i \tag{1}$$

Coefficients a_i are the result of analytical and statistical analysis of every individual trait. Every trait depends upon several dozen of handwriting characteristics. For them $a_i > 0$. For the rest characteristics, they equal 0. The procedure of evaluation of coefficients is not in the scope of the article.

Evaluation of x_i depends upon the definition of this handwriting characteristic. For instance, x_i for handwriting sign size reflects just the number of letters of a particular size category related to the total number of measured letters. Therefore, if in a specimen there are n inner letters of the middle zone and m of them belong to the big size, the corresponding $x_i = m/n$.

Expression (1) requires some refinement. First, not all handwriting signs can be evaluated at every text given for expertise. For instance, a small note on a piece of paper does not allow evaluation of margins. If a text is written on lined paper, we cannot evaluate inter-line distances. A specimen might not include diacritic signs or capital letters, thus we cannot evaluate the corresponding signs, etc. That is, the reliability of the trait evaluation depends upon the number of involved characteristics. Secondly, some traits in the model are presented through very few handwriting characteristics, others depend upon several dozen ones. It is clear that in the second case the model is more reliable. To consider this we are using an explicit reliability component. With this, the value of the modeled trait t depends upon calculated level y and reliability r (we are using a well-known Cobb-Douglas function for two factors):

$$t = y^{\alpha} \cdot r^{1-a} \tag{2}$$

Empirically we came to $\alpha = 0.6$. For r we are proposing the following expression:

$$r = 1\text{–}(1 - \mu)^k \tag{3}$$

Here μ is the reliability of the correct evaluation of the trait, assuming that we base the evaluation just on one handwriting characteristic. Therefore, 1-μ is actually the probability of an erroneous evaluation in this case. By $\mu = 0.2$, we come to the reliability value of 0.95 by 14 handwriting characteristics. Variable k is the number of handwriting characteristics.

The values for α and μ are empirical ones. That is quite acceptable since we are interested not in absolute values in the model (2), but rather in the possibility to formally compare different handwritings, i.e. different persons, to each other. We achieve that by means of the normalization of t. The HSDetect database contains hundreds of evaluated handwriting specimens. For normally distributed t

the normalized value (t^N) is calculated as followers:

$$t^N = \frac{t - T}{\sigma} \tag{4}$$

Since not all traits are normally distributed, the normalization in a general form can be calculated in the following way:

$$t^N = \frac{t - t^{min}}{t^{max} - t^{min}} \tag{5}$$

T is the average value of t, σ – standard deviation, t^{min} and t^{max} are correspondingly minimal and maximal values of t.

Model (1)–(5) is the most elaborated one. We call it E-Model. It is intended for particular psychological traits, where the number of independent variables, those for which $a_i > 0$ is relatively small and the differences in xi values are important. Examples of such traits are flexibility, ambition, arrogance, etc.

Relatively simple psychological constructs and test dimensions are presented with a bigger number of handwriting characteristics, e.g. 16 PF. This makes the usage of the reliability model and the quantitative presentation of individual handwriting characteristics less effective. The handwriting characteristics are modeled as binomial variables, that is, all xi take values 0 or 1. This model is called W-Model. It includes (1), (2), and (5), where $\alpha = 1$.

The plain P-Model is the further simplification of the W-Model, where all $a_i = 1$. That is, just the number of relevant handwriting characteristics, which are present in the specimen, defines the t value. P-Model is used for more complicated psychological constructs, like e.g. Big Five or some disease screening indicators.

Independent of the form any model must be thoroughly validated before it is practically implemented (Chernov, 2018; Chernov & Caspers, 2020). Below is an example of the Alzheimer's disease indicator.

2.2 *Alzheimer's Disease Screening Indicator*

Automatic writing is a learned skill. One of the leading American handwriting experts (Allen, 2016, p. 15) formulated it as follows: "Handwriting is a highly developed skill which we usually start to acquire in childhood and develop in the following years of adolescence and early adulthood. This is when handwriting becomes mature with an established form, barely changing over the years until factors such as illness and age start to impair it".

Handwriting requires cognitive work to retrieve the learned movements from the brain while at the same time compiling the text. Different diseases, which influence the brain and motoric functions, might affect handwriting at their earliest stages. Therefore, handwriting can effectively identify certain markers of dementia in

general and Alzheimer's disease (AD) in particular. Of course, handwriting analysis should not be seen as a diagnostic method. Diagnostics belongs to medicine. Handwriting analysis should be seen rather as an auxiliary instrument. However, in the forensic context, it can be very helpful.

Numerous studies confirm the strong relationship between AD and handwriting deterioration. So Croisile (Croisile, 1999) states: "Writing disorders are an early manifestation of Alzheimer's disease (AD), often more severe than language difficulties." Based on numerous studies by many researchers and the author's ownwa studies, the Alzheimer's Indicator (ADI) has been developed (Chernov & Zholdasova, 2021). It includes 40 handwriting and 2 linguistic characteristics. The formalization of this approach is defined by us as AD-HS instrument. Among the handwriting characteristics of AD-HS are different inconsistencies of handwriting (size, width, inter-letter, and inter-word intervals, slant, and pressure), big letter size, deterioration of letter forms and stroke quality, disconnected writing, specifics of diacritic and punctuation marks, and others. Linguistic characteristics include misspellings and obvious punctuation errors.

ADI is evaluated as P-Model, that is, the number of markers detected in the handwriting specimen of an investigated person divided by 42 (total number of characteristics). The subjects in our research were 47 participants, with whom AD was diagnosed by neurologists. They provided samples of their handwriting, mostly not written as part of the experiment, but rather their free writings. As a control group, 182 samples from the HSDetect database were taken.

The ADI value is calculated in the following steps. First step: an expert evaluates the analyzed handwriting and fills in the sign protocol, where all included handwriting and linguistic characteristics are listed. The protocol and the evaluation part were implemented for this experiment in Excel (all results are transferred also in the HSDetect database). To make the conclusion, whether a handwriting characteristic is present or not (that is enough for P-Model), he/she uses strict and unambiguous definitions of handwriting signs and characteristics, as it was stated above. As well above, we presented the example for the handwriting size. Second step: Excel calculates the number of present characteristics and divides it by 42 – the result is the ADI value.

The average ADI value for the control group was 0.26, with a standard deviation equal to 0.09. The average ADI value for the participants was 0.46. The comparison is shown in Fig. 1. The solid line (the ordinate dimension) denotes the ADI values of 47 subjects (their serial numbers are presented on the abscissa dimension), and the dashed lines from the center to the periphery denote correspondingly the average value, the 75% quantile, and the 95% quantile of the control group.

Besides, the correlation between ADI and the severity level of AD revealed by neurologists was 0.64 (based on 16 subjects). That is a high value and means that ADI can be effective for screening.

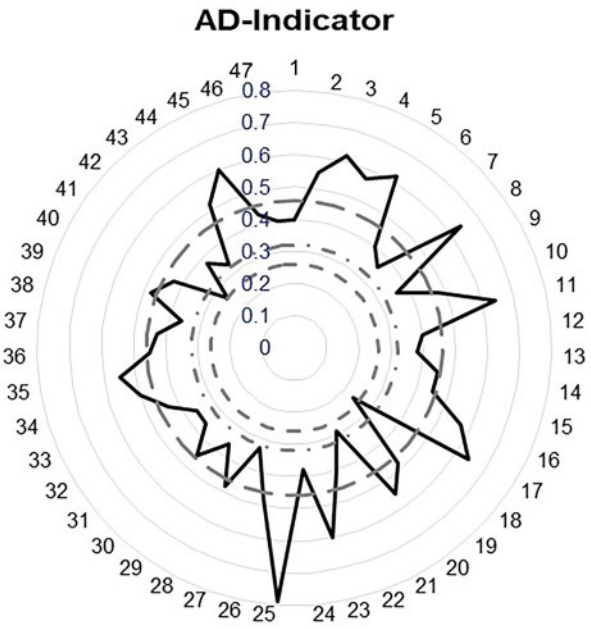


Fig. 1 Indicator of possible AD and its severity

2.3 Aggressivity Indicator

Aggressivity is a much more complicated construct than just aggressive behavior. Aggression can be defined as any act that harms another individual who is motivated to avoid such harm (Baron and Richardson 1994). This definition is very broad and covers a wide range of behaviors, starting from those without any harmful actions or passive-aggressive behaviors and ending with verbal and physical aggression that inflicts violence. It includes affective or reactive aggression, which is associated with a negative affect, typically anger, and instrumental aggression, which is typically goal-driven and could be free from affect. Aggressive actions are not always caused by the aggressiveness of the individual, and the aggressiveness is not always manifested in clearly aggressive actions. Therefore, modern studies emphasize the difference between the concepts of "aggression" and "aggressiveness". Aggressiveness is seen not only as the tendency of a person to act hostilely and aggressively, but as well as readiness for aggression. Aggressiveness is expressed in the conscious or unconscious aspiration of a person to cause someone or something harm, destroy, or damage it.

Aggressiveness is modeled in HSDetect by 31 handwriting characteristics (Chernov & Yengalychev, 2019). Among them are, for instance, angular connections, strong pressure, long in-stroke, tapering end-stroke, accented last letters in words, diminished upper zone, and others. The value of the aggressiveness construct

Table 1 Aggressiveness evaluation

Names	CL	JN	CM	CB	JD	JG	JV
AGI	0.52	0.52	0.71	0.39	0.35	0.52	0.35

(AGI) was calculated in the same way as in the previously described steps for the Alzheimer's indicator evaluation. In this case, also P-Model was used. Actually, we mostly use P-Model when the number of included handwriting signs is big (over two dozen). The average level of AGI in the HSDetect database equals 0.19 with a standard deviation of 0.09. To illustrate the model we evaluated the aggressiveness of seven famous American criminals,[4] in whose aggressiveness there is no doubt. The result is presented in Table 1.

In the first row, there are the initials of criminals, and in the second row – the value of the aggressiveness indicator. The values are much higher than with the "normal" people.

3 Discussion

The presented modeling of handwriting meets the standard requirements for a psychometric instrument: objectivity, reliability, and validity. Objectivity is ensured because, first, subjects cannot intentionally influence their handwriting. Secondly, the evaluation of handwriting signs is unambiguous and defined by formalized procedures. Thirdly, the evaluation of traits in HSDetect is computerized. The algorithms of evaluation are transparent; they implement the above-described models.

Reliability is ensured if experts by the building of sign protocol as the input for HSDetect follow the prescribed formal procedures of evaluation.

Validation should be provided for every individual trait. HSDetect database with hundreds of evaluated handwriting specimens is a proper reference base for that. Values can be normalized and compared.

With all mentioned advantages, the usage of handwriting psychology should not be overestimated. Handwriting is individual and unique, but we must not forget that many external factors can influence the analyzed specimen. Among them are paper, pen, external conditions (temperature and humidity), the changing culture of longhand, etc. Experts can consider them by their evaluation when they possess additional information about the specimen. However, often that is not the case. The best situation to overcome these difficulties is if we have several handwriting specimens of the same person. We just have to be very careful with the conclusions of the expertise. Nevertheless, formalized handwriting analysis can be of great help

[4] Charles Luciano (CL), John Hinckley (JN), Charles Manson (CM), Clyde Barrow (CB), John Dillinger (JD), John Gotti (JG), Joseph Valachi (JV).

for experts under special conditions, when other methods cannot be applied and can serve as an auxiliary instrument together with other psychometric methods.

References

Allen, M. (2016). *Foundations of forensic document analysis. Theory and practice*. Wiley-Blackwell.

Baron, R. A., & Richardson, D. (1994). *Human aggression*. Plenum.

Caligiuri, M. P., & Mohammed, L. A. (2012). *The neuroscience of handwriting*. CRC Press.

Chamorro-Premuzic, T., & Furnham, A. (2014). *The psychology of personnel selection* (3rd ed.). Cambridge University Press.

Chernov, Y. (2011). *сихологический анализ почерка: системный подход и компьютерная реализация в психологии, криминологии и судебной экспертиза* [Psychological analysis of handwriting: A systematic approach and computer implementation in psychology, criminology and forensics]. Genesis.

Chernov, Y. (2018). Formal validation of handwriting analysis. In Y. Chernov & M. A. Nauer (Eds.), *Handwriting research. Validation & quality* (pp. 37–68). Ebupli.

Chernov, Y. (2021). *Компьютерные методы анализа почерка* [Computer methods of handwriting analysis]. IHS Books.

Chernov, Y., & Caspers, C. (2020). Formalized computer-aided handwriting psychology: Validation and integration into psychological assessment. *Behavioral Sciences, 10*(1). https://doi.org/10.3390/bs10010027

Chernov, Y., & Yengalychev, V. (2019). Distant profiling: Aggression evaluation with formalized handwriting analysis. *Armenian Journal of Forensic Expertise and Criminalistics, 1*, 87–95.

Chernov, Y., & Zholdasova, Z. (2021). Markers of Alzheimer's disease in handwriting. *Russian Neurological Journal, 26*(6), 16–28. https://doi.org/10.30629/2658-7947-2021-26-6-16-28

Croisile, B. (1999). Agraphia in Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders, 10*, 226–230. https://doi.org/10.1159/000017124

Harralson, H. H., & Miller, L. S. (2018). *Huber and Headrick's handwriting identification. Facts and fundamentals*. Tailer & Franis.

Kanning, U. P. (2019). *Standards der Personaldiagnostik* [Standards of personal diagnostics] (2nd ed.). Hogrefe.

Michel, L. (1982). *Gerichtliche Schriftvergleichung. Eine Einführung in Grundlagen, Methoden und Praxis* [Forensic Handwriting Analysis. An Introduction to Basics, Methods and Practice]. Walter de Gruyter & Co.

Seibt, A. (1994). *Schriftpsychologie: Theorien, Forschungsergebnisse, wissenschaftstheoretische Grundlagen* [Handwriting Psychology: Theories, Research Results, Scinetific Baiscs]. Verlag Profil.

Analyzing Spatial Responses: A Comparison of IRT-Based Approaches

Amanda Luby, Thomas Daillak, and Sherry Huang

Abstract We investigate two approaches for analyzing spatial coordinate responses using models inspired by Item Response Theory (IRT). In the first, we use a two-stage approach to first construct a pseudo-response matrix using the spatial information and then apply standard IRT techniques to estimate proficiency and item parameters. In the second approach, we introduce the Spatial Error Model and use the spatial coordinates directly to infer information about the true locations and participant precision. As a motivating example, we use a study from forensic science designed to measure how fingerprint examiners use minutiae (small details in the fingerprint that form the basis for uniqueness) to come to an identification decision. The study found substantial participant variability, as different participants tend to focus on different areas of the image and some participants mark more minutiae than others. Using simulated data, we illustrate the relative strengths and weaknesses of each modeling approach, and demonstrate the advantages of modeling the spatial coordinates directly in the Spatial Error Model.

Keywords Bayesian statistics · Spatial statistics · Item response theory · Applications

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

A. Luby (✉) · T. Daillak · S. Huang
Department of Mathematics and Statistics, Swarthmore College, Swarthmore, PA, USA
e-mail: aluby1@swarthmore.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_31

1 Introduction

There are a number of tasks which may require individuals to identify spatial features on images including radiology, 'citizen science' image collection initiatives, epidemiology, and forensic science. There may be substantial variability in the number and location of coordinates that are marked by different individuals on the same image, which naturally lead to questions that Item Response Theory (IRT) is well-equipped to answer. For example, are some individuals stronger than others in identifying "true" features within the image? How much variability is attributable to individual differences, and how much is attributable to item differences (such as image quality)? While IRT exhibits potential to answer these questions, applying it to spatial data requires new techniques. IRT-based methods have been developed to incorporate spatial information (Santos-Fernandez & Mengersen, 2021; Cançado et al, 2016), but these approaches largely have a pre-defined set of (x, y) coordinates where responses occur. In cases where there can be any number of (x, y) coordinates marked by each individual, additional spatial considerations are needed.

Our aim is to apply IRT-based models for estimating participant proficiency in marking (x, y) coordinates when there is no ground truth for the expected responses. Such a model should allow for any number of coordinates to be marked on a given image. The motivating example for this work comes from forensic fingerprint analysis, where fingerprint examiners use *minutiae*, or small details in the fingerprint, to decide whether two prints came from the same source or not (See Fig. 1 for example). In studies designed to measure the accuracy and variability of fingerprint identification (Ulery et al., 2016b, 2014), multiple participants are asked to mark minutiae on a series of different fingerprints. In this setting, there is no ground truth for the number of minutiae or their (x, y) locations.

Two major national reports in the United States (PCAST, 2016; Council et al., 2009) have highlighted the subjectivity inherent to fingerprint analysis and the need

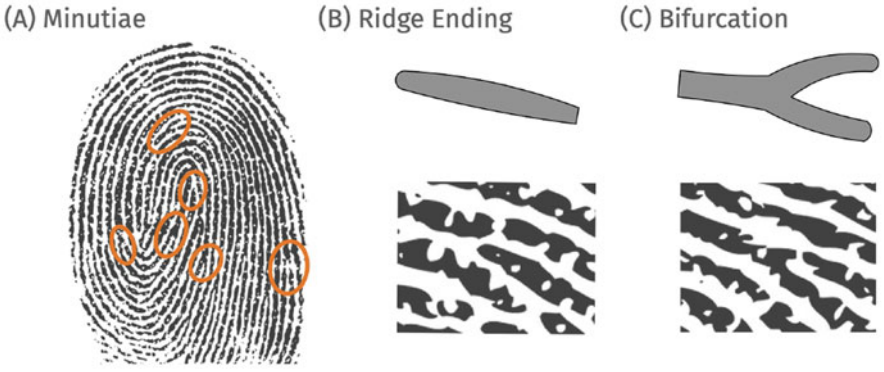


Fig. 1 Illustration of (**a**) a fingerprint with minutiae marked in orange, (**b**) a ridge ending, one type of minutiae, and (**c**) a bifurcation, a second type of minutiae

for research to quantify the variability in decision-making. These reports led to a variety of 'black box' error rate studies, e.g. Ulery et al. (2011), Pacheco et al. (2014), designed to estimate error rates in realistic casework. Ulery et al. (2016b) and Ulery et al. (2014) instead treated examiners as 'white box' decision-makers and measured the factors used to come to a final decision. Since minutiae identification is a key part of the decision-making process, individual differences could have massive downstream impacts on the final results. We would like to be able to use IRT-like machinery to quantify performance of individuals in these minutiae identification tasks.

2 Background

2.1 White Box Study

As a motivating example, we use results from the FBI "White Box" study (Ulery et al., 2016b). One-hundred and seventy fingerprint examiners were recruited to participate. Participants were shown a subset (generally 20–25 per examiner) of 320 fingerprint pairs, intentionally chosen to represent a broad range of quality including low-quality fingerprints that would likely not progress through casework. Participants were first shown the *latent print*, the image of a fingermark taken from a crime scene, and were asked to select minutiae and provide a quality assessment. Next, they were shown the *reference print*, a high-quality, known-source image taken under idealized conditions, and asked to do the same minutiae identification task. After marking both the latent and reference print, they were shown the two marked up images together and were asked to compare the two and come to a final decision. In this stage, they were allowed to add, delete, or move minutiae on either print.

Because of the complexity of the task and variety of skill sets involved, we focus solely on the *analysis phase* of the latent print. This allows us to isolate a single latent trait of the examiners: their proficiency at marking minutiae on a latent print. For further details on the fingerprint examination process and current recommendations and practices, see Friction Ridge Subcommittee of the Organization of Scientific Area Committees for Forensic Science (2017, 2019).

Figure 2 shows the total number of minutiae marked by each examiner throughout the entire study, along with whether each minutiae was classified as a *supermajority* (minutiae was marked by $> \frac{2}{3}$ of participants who were assigned the image), a *majority* (minutiae was marked by $> \frac{1}{2}$ of assigned participants), a *minority* (More than 1, but fewer than $\frac{1}{2}$ of participants marked it), or a *singleton* (that minutiae was marked by only one participant). Singletons make up the majority of the marked minutiae for every examiner in the study. Figure 3 shows the percentage of classified minutiae for each examiner after excluding minutiae that were only marked by one or two participants. This allows us to visualize the

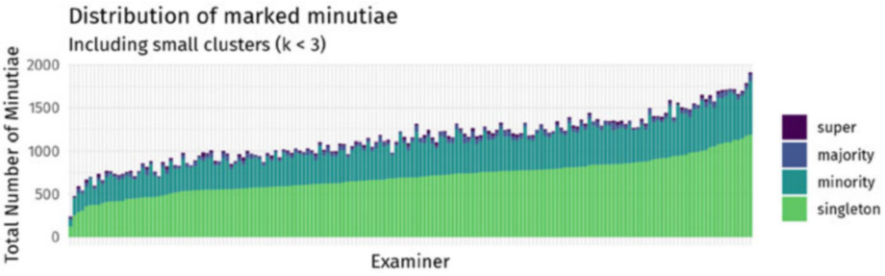


Fig. 2 Number of marked minutiae and their classifications for each participating examiner. Singletons are by far the most common, and there is a large discrepancy in the total number of marked minutiae across examiners

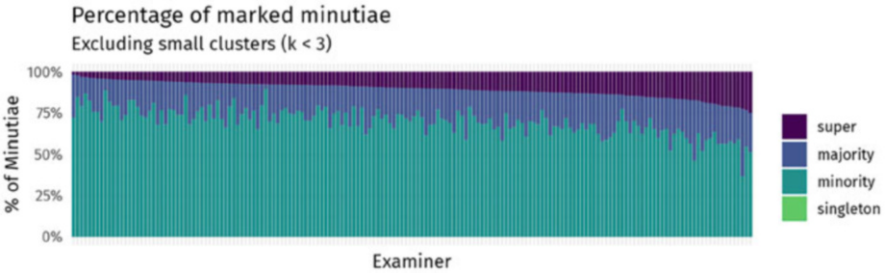


Fig. 3 Percentage of marked minutiae classified as supermajority, majority, and minority for each participating examiner (excluding clusters of size 2 or fewer)

variability in how often participants agree with the majority, even after excluding the highest variability markings. Taken together, these figures demonstrate substantial variability in examiner markup, but examiners cannot be directly compared to one another, since each examiner was assigned a different subset of fingerprint images.

3 Methods

We will compare two approaches for analyzing this type of data: (1) a two-stage procedure that firsts clusters the data according to the original study to create a pseudo item response matrix and then applies a Rasch model (Rasch, 1960; Fischer and Molenaar, 2012), and (2) an IRT-like model that uses the (x, y) coordinates directly as a response that we call the *Spatial Rater Model*.

3.1 Two-Stage Rasch Model

In Ulery et al. (2016a), the marked minutiae were clustered using the DBSCAN algorithm (Hahsler et al., 2019), followed by hierarchical clustering to split over-

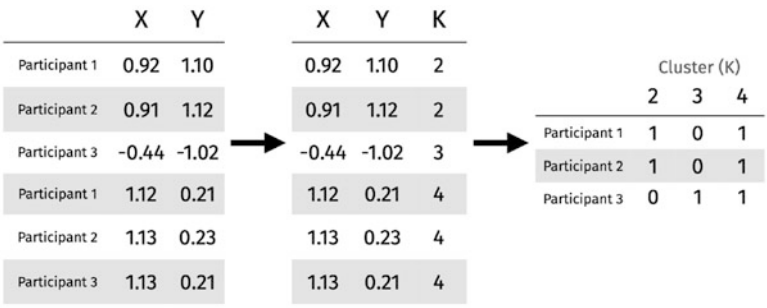


Fig. 4 An example of the creation of cluster labels for each recorded (x, y) coordinate followed by the creation of the pseudo item response matrix

large clusters (Anderberg, 2014). This results in each $(x, y) \in \boldsymbol{X}$ being assigned a cluster label $k = 1, 2, \ldots, K$. After cluster labels have been determined, each cluster can be treated as a pseudo-item. This results in a $I \times K$ pseudo item response matrix, $\boldsymbol{Y}$, where $Y_{ik} = 1$ if examiner i marked a minutiae in cluster k and 0 otherwise. The clustering approach allows us to pool information across different fingerprints, since not every examiner was shown every fingerprint. A visual representation of constructing the pseudo-response matrix is shown in Fig. 4.

Once $\boldsymbol{Y}$, the pseudo-response matrix, has been constructed, a standard Rasch model can be used to estimate participant (θ_i) and cluster (b_k) parameters:

$$P(Y_{ik} = 1) = \text{logit}^{-1}(\theta_i + b_k). \tag{1}$$

The latent scale of this Rasch model is different than the usual latent proficiency scale of a Rasch model, and represents a 'liberal to conservative' scale in marking minutiae. We expect that there are some examiners who mark liberally, and others who mark more conservatively (see, e.g., Figs. 2 and 3). Since we have not applied any type of scoring to the marking of clusters, larger θ values will indicate examiners who mark more liberally, and smaller θ values will indicate examiners who mark more conservatively. Likewise, higher b values will indicate minutiae that were marked by more participants, and smaller b values will indicate minutiae that were marked by fewer participants.

However, a more concerning problem is that this model ignores any systematic error in the location that minutiae are marked, and so it is impossible to distinguish participants who are more precise (i.e. closer to the true location) from participants who are less precise.

3.2 Spatial Rater Model

The second analysis approach attempts to correct for the weakness identified above by modeling the marked (x, y) coordinates explicitly. We assume that each marked

minutiae, $\boldsymbol{X_{ik}}$, corresponds to a ground truth minutiae location $\boldsymbol{T_k}$, that is located with some error (ϵ_{ik}) that depends on participant spatial competency, θ_i. Participants with strong spatial competency (corresponding to more negative θ values) should generally mark X_{ik} that are very close to T_k (small ϵ_{ik}). Participants with weaker spatial competency will generally mark X_{ik} that are further away from T_k (larger ϵ_{ik}). The hierarchical Bayesian model specification is as follows:

$$
\begin{aligned}
\boldsymbol{T_k} &\sim MVN(\mathbf{0}, \Sigma) && \text{True minutiae locations (scaled)} \\
\boldsymbol{X_{ik}} &= \boldsymbol{T_k} + \epsilon_{ik} && \text{Respondent i's appraisal of minutiae k} \\
\epsilon_{ik} &\sim \text{lognormal}(\theta_i, 1) && \text{Error depends on participant competency} \\
\theta_i &\sim N(\mu_\theta, \sigma_\theta) && \text{Participant competencies are normally distributed} \\
\Sigma &= \sigma_T L_T L_T' \sigma_T && \text{Cholesky decomposition is used for efficiency} \\
\sigma_T &\sim \text{Half-Cauchy}(0, 5) && \text{Weakly informative prior} \\
L_T &\sim LKJ(2) && \text{Correlation mildly concentrates around the identity}
\end{aligned}
$$

We assume that the true minutiae locations ($\boldsymbol{T_k}$) follow a multivariate normal distribution. Since minutiae locations must have an x and y coordinate, this allows us to model them jointly. It also assumes that minutiae are more likely near the core of fingerprint (assumed to be $(0, 0)$). We assume that the error, ϵ_{ik}, is log-normally distributed. This ensures that the error is positive, so error in all directions is treated equivalently. In the future, it may desirable to, e.g., model the error in the x-direction different than the y-direction, in which case a different prior distribution should be chosen.

There are a few notable advantages of the Spatial Rater Model as compared to the two-stage Rasch model. First, the ground truth locations, $\boldsymbol{T_k}$, need not be defined a priori but are estimated by the model. Second, the θ parameter corresponds to how precise an individual is at marking minutiae. Smaller θ values correspond to higher precision (and smaller error) while larger θ values correspond to less precision (and larger error). However, no additional pseudo-item parameters are estimated beyond the locations, and so the model does not account for the possibility that some minutiae are harder to locate than other minutiae.

4 Results

The images of the fingerprints themselves are not available, and so we cannot verify the true minutiae locations. We instead use simulated data to assess the performance of each modeling approach.

4.1 Simulated Data

In order to evaluate the two modeling approaches, we must first simulate data for which the true participant parameters and minutiae locations are known. The following simulation procedure was used.

1. Fit the Spatial Rater Model (as defined in Sect. 3.2) to the White Box dataset to find a reasonable distribution for $\boldsymbol{T_k}$. In our case,

$$\boldsymbol{T_k} \sim MVN(\mathbf{0}, \begin{bmatrix} 0.90 & 0.07 \\ 0.07 & 1.06 \end{bmatrix})$$

2. Simulate 20 minutiae ($T_1, T_2, \ldots, T_K$, $K = 20$) for each of 10 items ($J = 10$).
3. Draw $\theta_i \sim N(-7, 3)$ for $i = 1, \ldots, 100$ participants (The distribution parameters can be chosen from the results of the fit in Step 1).
4. Simulate participant i's location of minutiae k, X_{ik}, according to $X_{ik} = \boldsymbol{T_k} + \epsilon_{ik}$, where $\epsilon_{ik} \sim \text{lognormal}(\theta_i, 1)$.

Following steps 1–4 of this simulation procedure results in minutiae locations that are fully observed, where every participant has marked every minutiae. As shown in Fig. 2, this scenario is far from realistic in practice. To create a more realistic dataset, an additional parameter for each individual was simulated (π_i) to determine the probability that they mark any given minutiae.

5. For each participant, draw $\pi_i \sim \text{Unif}(0.1, 0.4)$
6. For each participant $\times$ minutiae pair, draw $Z_{ik} \sim \text{Bernoulli}(\pi_i)$
7. Define

$$X'_{ik} = \begin{cases} X_{ik} & Z_{ik} = 1 \\ X_{ik} = NA & Z_{ik} = 0 \end{cases}$$

and use X'_{ik} as the simulated responses.

The resulting true minutiae locations ($\boldsymbol{T_k}$) for each simulated fingerprint, along with the corresponding simulated marked minutiae, X'_{ik}, are shown in Fig. 5. While the marked minutiae are generally close to the true minutiae locations, some are further away and there is considerable overlap in certain areas (e.g. near the center of Simulated Fingerprint 8).

4.2 Two-Stage Approach

Following the model outlined in Sect. 3.1, we first cluster the minutiae locations from each participant in order to create a pseudo item response matrix $\boldsymbol{Y}$ (as displayed in Fig. 4). In this matrix, each row corresponds to a (simulated) participant

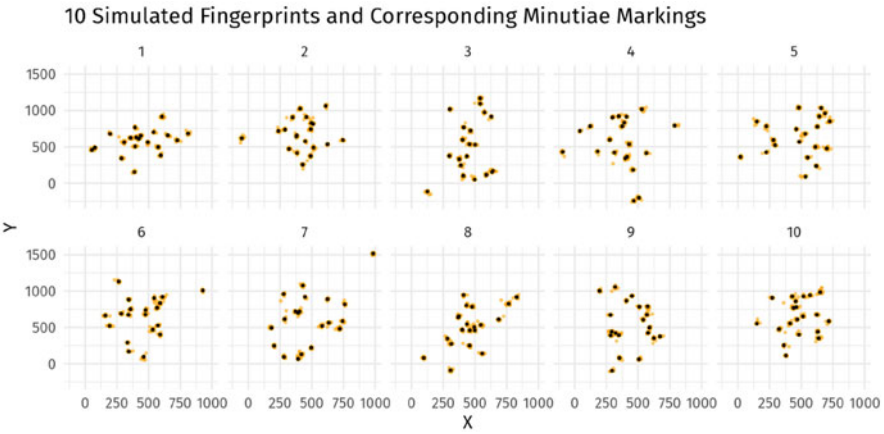


Fig. 5 Simulated fingerprints (minutiae shown in black) and corresponding simulated minutiae markup from 100 examiners (in orange)

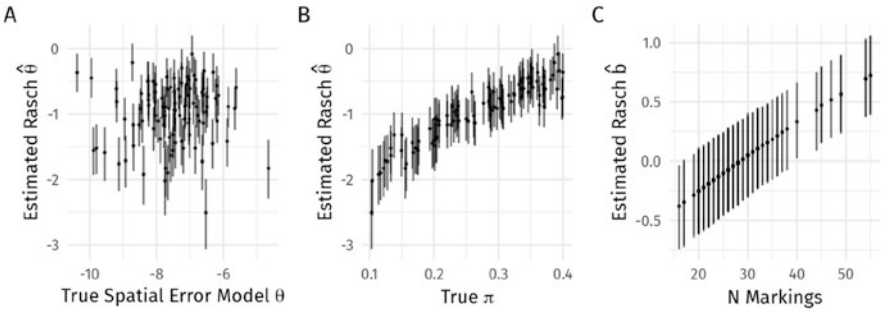


Fig. 6 Estimated parameters from the two-stage Rasch model on the simulated data. Plots (**a**) and (**b**) display the $\hat{\theta}$ estimates, with 95% posterior intervals, compared to the true Spatial Error Model θ and π values used to simulate the minutiae markings. Plot (**c**) displays the $\hat{b}$ estimates, with 95% posterior intervals, compared to the total number of participants (out of 100) that marked each of the pseudo-items

and each column corresponds to a cluster. After the response matrix $\boldsymbol{Y}$ has been constructed, a Rasch model is fit in order to estimate participant 'proficiency' and cluster 'difficulty'. However, the data generating process outlined in Sect. 4.1 does not rely on proficiency of individual participants, or difficulty of particular minutiae, since the data were generated with only spatial error for each participant. Instead, the estimated θ represents a tendency to mark more or less minutiae.

Figure 6a shows the estimated proficiencies from the Rasch model compared to the true spatial competency of each participant. As expected, there is no clear relationship between these two quantities. The two-stage approach is unable to recover the true spatial competencies of each participant, since the (x, y) coordinates were transformed into a pseudo-response matrix. Instead, the Rasch θ's correspond to

how likely an individual is to mark any given minutiae, which is governed in the simulation by the π parameter. This relationship is shown in Fig. 6b and shows a strong trend. Finally, Fig. 6c shows the estimated cluster b_k's from the Rasch model, which are determined by how often each cluster was marked in the simulation. Clusters that were marked more often were assigned larger b_k's than clusters that were marked less often.

These results underscore the weakness of using the two-stage approach. It is impossible to recover the spatial competencies of each participant, and the θ and b parameters estimated by the Rasch model are both driven by participant propensity to mark additional minutiae (π).

4.3 Spatial Rater Model

Finally, we fit the *Spatial Rater Model* defined in Sect. 3.2 on the simulated data. Since the data were generated according to this model, the parameters should be able to be recovered. Figure 7 shows the true minutiae locations, $\boldsymbol{T_k}$, in black and the detected minutiae in green. In general, the model is able to recover the true locations well, but misses some true minutiae in high-density areas (e.g. the center of Fingerprints 3 and 8) or minutiae that are close to others (e.g. Fingerprints 4 and 10).

Next, we investigate whether the estimated spatial competencies ($\hat{\theta}$) recover the true spatial competency values (θ) in Fig. 8a, and if there is any relationship to the tendency to mark minutiae (π) in Fig. 8b. While Fig. 8a shows a very strong correlation between $\hat{\theta}$ and θ, the model does tend to systematically under-estimate

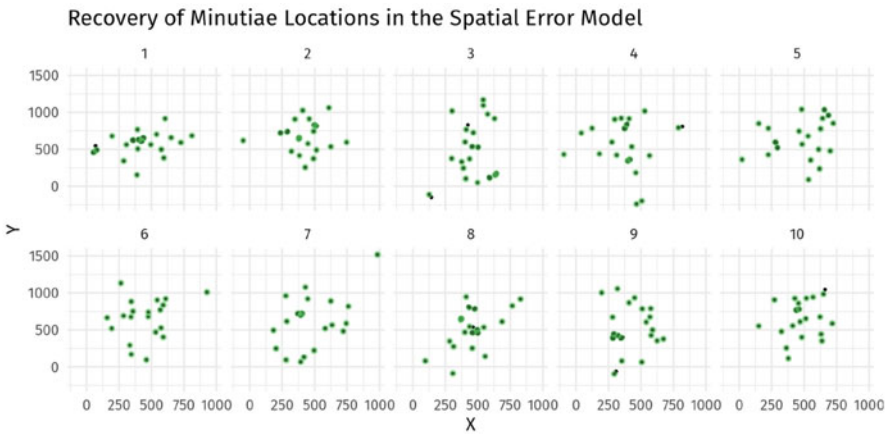


Fig. 7 Recovery of minutiae locations (T_k) in the Spatial Error Model. True minutiae locations are shown in black, and estimated T_k are shown in green. The model generally recovers the true T_k well, but misses some true minutiae in high-density areas

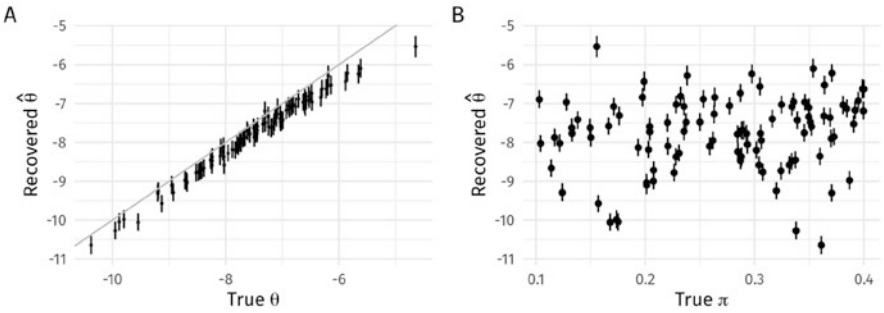


Fig. 8 Parameter recovery in the Spatial Error Model. The θ estimates are strongly correlated with the true values, but the model tends to under-estimate the spatial competency of individuals. There is no discernible relationship between $\hat{\theta}$ and π

the θ values, likely due to shrinkage from the prior distribution. As expected, there is no discernible relationship between $\hat{\theta}$ and π. Overall, the model shows promise for modeling error in spatial responses when the number of marked spatial features per item may vary, but further study and refinement is needed prior to implementation in practice.

5 Discussion

Measuring individual performance in identifying spatial coordinates could provide novel and valuable information about variability in decision-making in a variety of settings. Using fingerprint analysis as a motivating example, we demonstrate the use of a two-stage procedure with standard IRT machinery on simulated data, and introduce the Spatial Error Model as an alternative that models the spatial responses directly. Overall, parameter recovery of the Spatial Error Model was strong and provided improvements over the two-stage approach, which required some loss of information.

However, the results presented here represent a single simulation, and correspond to one set of chosen parameter values. A natural next step of this work is to evaluate the model on further replications of these parameter values, as well as additional sets of parameter values and data-generating processes. The primary consideration of this initial model was also to recover the true minutiae locations, and future iterations of the model will focus on improvements to the interpretability of the participant spatial competencies and associated errors. The model must be shown to be consistent and robust prior to implementation in practice, which requires additional simulation studies and model refinement.

Additionally, the algorithm used to initially cluster the (x, y) coordinates (DBSCAN, Hahsler et al., 2019) is sensitive to small changes in the ϵ parameter. Since we do not have access to the original images, there is no way to independently

perform a sensitivity analysis on the original data and resulting cluster labels. A further area of future work is varying this parameter on the simulated data to determine how sensitive the resulting latent parameter estimates are to such changes.

Both modeling approaches discussed in the current paper rely on the initial clustering step in order to determine which (x, y) coordinates marked by different individuals correspond to the same minutiae. This introduces a "double-dipping" problem (Kriegeskorte et al., 2009), where the data are used twice: once to construct the clusters and once to fit the model. This can result in artificially small uncertainty estimates and uncontrolled Type I error. A future area of expansion is correcting for this through incorporating the clustering step in a hierarchical model.

References

Anderberg, M. R. (2014). *Cluster analysis for applications: Probability and mathematical statistics: A series of monographs and textbooks* (Vol. 19). Academic Press.

Cançado, A. L., Gomes, A. E., da Silva, C. Q., Oliveira, F. L., & Duczmal, L. H. (2016). An Item Response Theory approach to spatial cluster estimation and visualization. *Environmental and Ecological Statistics, 23*(3), 435–451.

Council, N. R., et al. (2009). *Strengthening forensic science in the United States: A path forward*. National Academies Press.

Fischer, G. H., & Molenaar, I. W. (2012). *Rasch models: Foundations, recent developments, and applications*. New York: Springer Science & Business Media.

Friction Ridge Subcommittee of the Organization of Scientific Area Committees for Forensic Science. (2017). *Guideline for the articulation of the decision-making process leading to an expert opinion of source identification in friction ridge examinations*. Online; Accessed 15 September 2021.

Friction Ridge Subcommittee of the Organization of Scientific Area Committees for Forensic Science. (2019). *Friction ridge process map (current practice)*. Online; Accessed 15 September 2021.

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast density-based clustering with R. *Journal of Statistical Software, 91*(1), 1–30. https://doi.org/10.18637/jss.v091.i01. https://www.jstatsoft.org/index.php/jss/article/view/v091i01

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience, 12*(5), 535–540.

Pacheco, I., Cerchiai, B., & Stoiloff, S. (2014). *Miami-Dade research study for the reliability of the ACE-V process: Accuracy & precision in latent fingerprint examinations*. Unpublished report (pp. 2–5).

PCAST, P. (2016). *Report to the President, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-comparison Methods*. Executive Office of the President of the United States, President's Council ...

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Santos-Fernandez, E., & Mengersen, K. (2021). Understanding the reliability of citizen science observational data using Item Response Models. *Methods in Ecology and Evolution, 12*(8), 1533–1548.

Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2011). Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences, 108*(19), 7733–7738.

Ulery, B. T., Hicklin, R. A., Roberts, M. A., & Buscaglia, J. (2014). Measuring what latent fingerprint examiners consider sufficient information for individualization determinations. *PloS One, 9*(11), e110, 179.

Ulery, B. T., Hicklin, R. A., Roberts, M. A., & Buscaglia, J. (2016a). Data on the interexaminer variation of minutia markup on latent fingerprints. *Data in Brief, 8*, 158–190.

Ulery, B. T., Hicklin, R. A., Roberts, M. A., & Buscaglia, J. (2016b). Interexaminer variation of minutia markup on latent fingerprints. *Forensic Science International, 264*, 89–99.

Application of the Network Psychometric Framework to Measurement Burst Designs

Michela Zambelli, Semira Tagliabue, and Giulio Costantini

Abstract Network Psychometrics emerged in the last years as an approach that allows investigating how different elements of a system interact and how these interactions change across occasions. The present work aims to show the potentialities of the Network Psychometric framework to examine the stability of dynamics of change of psychological processes. Specifically, we tested the applicability of the recently introduced *psychonetrics* toolbox to (a) model within-subjects (both contemporaneous and temporal) and between-subject (stable individual differences) dynamics with data collected with a measurement burst design (MBD, two 14-day bursts); and (b) examine the temporal stability (or instability) of the process' dynamics by directly comparing the two bursts in terms of both within and between parameters' invariance. The illustrative example was about the process of meaning-making, whose dynamics of change were examined across two different contextual conditions during the COVID-19 pandemic. A step-by-step procedure to apply *psychonetrics* to MBDs is provided in an Open Science Framework project.

Keywords Network psychometric · Psychonetrics · Measurement burst design · Meaning-making

M. Zambelli (✉)
Università di Genova, Genova, IT, Italy

Università Cattolica del Sacro Cuore, Milano, IT, Italy
e-mail: michela.zambelli@unicatt.it

S. Tagliabue
Università Cattolica del Sacro Cuore, Brescia, IT, Italy
e-mail: semira.tagliabue@unicatt.it

G. Costantini
University of Milan Bicocca, Milano, IT, Italy
e-mail: giulio.costantini@unimib.it

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_32

1 Introduction

The network psychometric approach offers the best representation of the concept of dynamic system of interacting elements (Borsboom et al., 2021). This approach emerged in the last decade as an alternative way to the traditional latent variable approach to investigate patterns of associations among variables in a multivariate framework (Borsboom & Cramer, 2013). Based on the idea of the mutualism model (van der Maas et al., 2006), each psychological process (e.g., intelligence) is conceptualized as a complex system made of several elements (i.e., memory, decision, reasoning) in a dynamical interaction, from which the development of the entire system is generated (Borsboom et al., 2021; Marsman & Rhemtulla, 2022). Within this framework, a psychological process can therefore be visualized as a network of intercorrelation ("edges") between the constitutive elements ("nodes") of the system (Borsboom et al., 2021). A cross-sectional network is usually estimated modelling observed indicators as the nodes of the system and the connections between these nodes as partial correlations, that are unique associations between each couple of nodes after controlling for the associations with all the other nodes (Epskamp et al., 2018).

1.1 Psychonetrics: A Toolbox for Confirmatory Testing in Network Psychometrics

Thanks to its data driven approach, the network analysis methodology was introduced as a powerful tool for exploratory research (Epskamp et al., 2018), to be used when prior knowledge about process dynamics is not sufficient to make strong causal hypothesis. However, recently researchers dealt with the challenge of extending the network psychometrics to test confirmatory hypothesies, for instance evaluating group differences in the network structure (Marsman & Rhemtulla, 2022). To fill this gap, a new toolbox was developed named *Psychonetrics* (Epskamp, 2020a). The *psychonetrics* toolbox allows combining the exploratory search of the Gaussian Network Modeling with the Confirmatory testing of the SEM framework, by introducing fit indices, parameter significance and the possibility to evaluate group differences in the network structure by comparing nested models (http://psychonetrics.org/).

Recently, network models from time-series and panel data have been developed to offer a thoughtful insight into multivariate pattern of temporal dynamics of psychological processes collected from multiple individuals (Borsboom et al., 2021; Epskamp, 2020b). Multilevel temporal networks (when time is nested within people) can be estimated as *mlVAR* models that possess three basic assumptions to be verified before running models: the normality of item distribution, the stationarity of parameters, and the equality of time intervals (Epskamp et al., 2018). Currently, despite several modelling techniques for intensive and longitudinal data have been implemented in the *psychonetrics* package, there is a lack of empirical studies

testing the applicability of its confirmatory approach to real data collected with intensive longitudinal designs.

This work presents an application of the *ml_ts-lvgvar* model (Epskamp, 2020b), that combines the multilevel graphical vector-autoregression (*mlGVAR*) framework with the latent variable modelling (*lvm*) for time series data with a nested structure (days nested within person). The model is estimated following the variance decomposition in a within-level variance, encoding dynamic effects, and a between-level variance, representing individual differences (Epskamp, 2020b). At the within-level, two networks are generated: the *temporal network* is generated from a matrix of directed vector autoregressive coefficients to assess if a deviation from a subject's mean predicted a deviation from a subject's mean in the same component (i.e., inertia) or in another component (i.e., temporal influence) at the next measurement occasion; the residual matrix is further modelled as a *contemporaneous network* mapping the within-person partial correlations (i.e., concurrent associations) between the components within the same day, after conditioning for the previous measurement occasion. A third matrix can be estimated to form a GGM (gaussian graphical model) encoding how the stationary means of different subjects relate to one another, this is called *between-person network*.

In the present study we applied the *multilevel ts-lvgvar* model to real data collected with a measurement burst design made of two 14-days diary studies. The multi-group function available for this model in *psychonetrics* was used to directly compare the dynamic structure of the two waves and infer about the stability of process dynamics across time.

2 Illustrative Example

The illustrative example is dedicated to examining the stability of dynamics of change of the meaning-making process, defined as the process by which individuals build the meaning of their life (Park et al., 2012; Steger et al., 2009). The meaning-making process has been conceptualized as a system of interacting elements made of six basic components (presence of meaning, presence of significance, presence of purpose, search for presence, search for significance, search for purpose; Zambelli & Tagliabue, 2023).

2.1 Method

Data was collected with a measurement burst design composed of two 14-days daily diary studies from a sample of emerging and young adults (18–35 years; Arnett, 2007). The first burst occurred in March 2020 during the first Italian lock-down, the second burst was 10 months later (February 2021), when the restrictions imposed due to the pandemic were temporally eased. Participants signed an informed consent before participating in both waves. Ethical approval was issued by the Ethics

Committee of Università Cattolica del Sacro Cuore of Milan (Italy). Among the 529 participants (M_{age} = 25.5; SD = 4.1; Males = 27%; Students = 44%), the 27.6% (N = 146) took part in both waves, the 34.4% (N = 182) completed only the first wave and the 38% (N = 201) only the second wave. Participants completed on average 12.6 daily questionnaires (range = 1–14; SD = 2.9) in Wave 1, and 11 questionnaires (range = 1–14; SD = 4.1) in Wave 2.

In both waves, participants completed a short online questionnaire at 7 p.m. for 14 consecutive days, including the SMILE measure (Zambelli & Tagliabue, 2023). Following Nezlek (2017) indications, an unconditional two-level model (days nested within person) was conducted to extract multilevel descriptive statistics of the six items (mean and within- and between-person variance) in the two waves. Descriptive statistics and mean-level comparisons across the two waves are presented in Table 1.

Table 1 Multilevel descriptive statistics of the SMILE items and comparison across the two waves

	Variance						
	Within-level		Between-level		Grand-mean		Grand-mean comparison
SMILE – items	W1	W2	W1	W2	W1	W2	t(df), *p* value
Presence of comprehension Today, I think I comprehend the meaning of my life during this pandemic	1.20	1.07	2.27	2.09	3.46	3.61	24.96 (7848), <.001
Presence of significance Today, I feel that my life has value during this pandemic	1.22	1.23	2.26	1.95	4.04	4.04	7.68 (7852), .89
Presence of purpose Today, I think I have goals for my life that push me to move forward during this pandemic	1.20	1.31	2.15	1.85	4.25	4.14	7.80 (7855), .01
Search for comprehension Today, I tried to understand the meaning of my life during this pandemic	1.14	1.10	1.56	1.60	2.61	2.81	3.44 (7847), <.001
Search for significance Today, I tried to understand what values my life in this pandemic	1.39	1.21	1.67	1.62	2.90	2.96	17.21 (7854), .08
Search for purpose Today, I searched goals for my life that will push me to move forward during this pandemic	1.59	1.35	1.74	1.59	3.11	3.10	31.64 (7850), .77

W1: wave 1, *W2*: wave 2, *t*: statistical value of Student's t-test. Items were rated on a 7-point Likert scale

2.2 Analytic Strategy

Statistical analyses were conducted with the R Version 4.1.2. Codes are available at: https://osf.io/bsmhn/?view_only=d8395f62a25e4eed8e6e54de8ee1a26e

Step 1 – Verify Model's Assumptions Before running the multilevel network model, it is necessary to verify that the three basic assumptions of VAR (vector auto-regressive) models are respected in the data. These are: the normality of items distribution; the *stationarity* of temporal dynamics (i.e., dynamic parameters are constant over time); and the equality of time-intervals (i.e., measurement occasions are equally spaced). The normality of distribution was examined by verifying that items' values of kurtosis and skewness did not exceed $\left| 1.2 \right|$ (Muthén & Kaplan, 1985). To examine the stationarity of the process, we fitted a series of fixed-effects linear regressions (with alpha set to .05) with the day number as predictor to each of the six variables of the SMILE to check for any linear *trends* over time, represented by a systematical increase or decrease of values across measurement occasions. If trends were present, data were detrended following the procedure presented in Borsboom et al. (2021). Finally, the equality of time intervals was guaranteed by the research design as data were collected every 24 hours.

Step 2 – Estimate a Multilevel Network Model To answer the first aim, we examined data from the first wave. The multilevel network model that we planned to apply required the dataset to be set in long format (each row indicated one person at one time point). A *ml_ts-lvgvar* model was estimated using version 0.10 of the *psychonetrics* package. The six detrended SMILE items were included as the nodes of the network and missing data were handled with the FIML (full information maximum likelihood) estimator (Epskamp, 2020b). The adaptability of the model to our data was evaluated through fit indices provided by the *psychonetrics* R package: the χ^2 (Cheung & Rensvold, 2002), the comparative fit index (CFI; acceptable fit for values $\geq$.90; Little, 2013), the root mean square error of approximation (RMSEA; acceptable fit for values $\leq$.08; Little, 2013), the AIC (Akaike Information Criterion; Akaike, 1987) and the BIC (Bayesian Information Criterion; Schwarz, 1978) for which lower values are desirable. After that, we estimated three network structures from the respective matrices (contemporaneous, temporal and between-person), and we visualized the networks with the *qgraph* package (Epskamp et al., 2012).

Step 3 – Estimate a Multi-group Multilevel Network Model A multi-group *ml_ts_lvgvar* model was estimated using the *groups* function available in the *psychonetrics* package by indicating as grouping variable a dummy variable indicating the belongingness to wave 1 or wave 2. In this study the two waves of the measurement burst design were considered as independent samples, in the discussion section the implications of this choice are discussed. In this model, the parameters of the three matrices (contemporaneous, temporal and between) were free to vary across the two levels of the grouping variable. Then, we fitted three constrained models, in which the parameters of the three matrices were forced to equality across the two groups one by one, by using the *groupequal* function. The

fit indexes of the constrained models were compared with the free model to evaluate the existence of any significant differences. We relied on the BIC (Schwarz, 1978) to identify significant differences between the nested models as it is the most restrictive index that penalizes model complexity. The BIC weights the estimate according to the degrees of freedom of the model as indicated by the following equation: BIC = T − df ln(N), where T is the chi-square test statistic of the model, df are the degrees of freedom of the model, and N is the number of cases (Lin et al., 2017).

2.3 Results

Participants with less than 80% of missing values in the SMILE items within each wave were retained for a total of 318 cases in Wave 1 and 320 case in Wave 2. The VAR model's assumptions were verified in both waves. To obtain model convergence, a standardization with s-scores was required.

Investigating the Dynamics of the Meaning-Making Process in a Daily Framework The overall model, conducted on the first wave, showed acceptable fit indices ($\chi^2(3570) = 32670.5$, $p < .001$; CFI = .89; RMSEA = .053 [.051–.055]; AIC = 43859.7; BIC = 44175.7). The three network structures are represented in the wave 1 section in Fig. 1. Since the aim of the current paper is to elucidate the application of the multilevel network psychometric, below we report the interpretation of the main effects to clarify what kind of information on the process dynamics can be obtained from each network.

Contemporaneous Network From the graphical visualization two clusters of nodes are visible; first cluster encloses the three items of presence of meaning, the second cluster is formed by the three items of search for meaning. Within each cluster, the nodes were connected by thick lines, with partial correlations always above .20.

Temporal Network All the nodes showed a significant autoregressive effect (self-loops), thus suggesting the presence of a trait dimension for each meaning-making features that is rather stable over time. Each node has a specific role within the network; some nodes are more central as they share lot of connections with other nodes (e.g., search for purpose; MILSP) while other nodes have a marginal role in the dynamics of the system (e.g., presence of purpose; MILPP). Some bi-directional temporal associations (directional arrows between couples of nodes) are also present. For instance, presence of comprehension (MILPC) and presence of significance (MILPS), such that an increase in presence of comprehension on one day predicted an increase in presence of significance the day after and vice-versa.

Between Network This network includes the partial correlations between the mean levels of nodes across the 14 days. The three items belonging to the same cluster were positively correlated, meaning that, over the 14 days of the first wave, people experienced similar levels in the three presence of meaning components, and similar

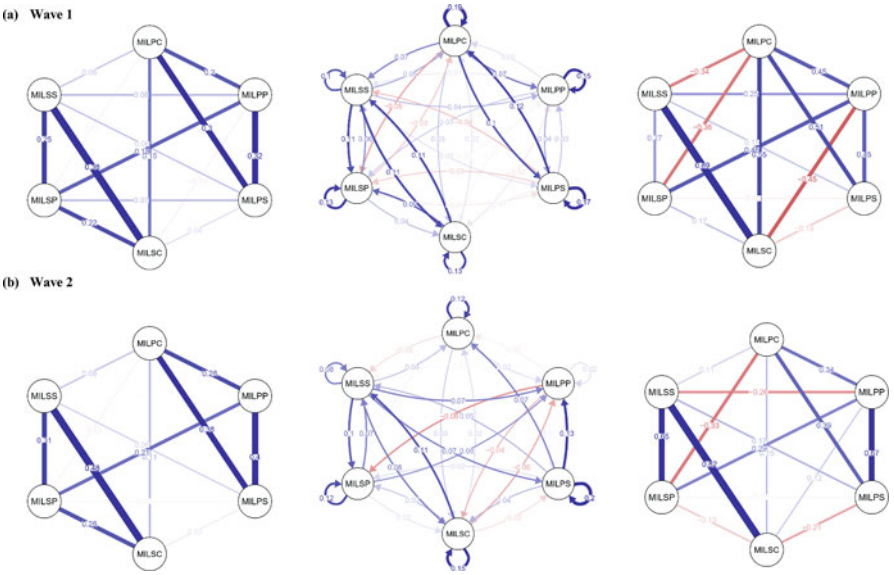


Fig. 1 Representation of the contemporaneous, temporal and between-person network across the two waves *Note.* Left: contemporaneous network; center: temporal network; right: between-person network. *MILPC*=presence of comprehension, *MILPS*=presence of significance, *MILPP*=presence of purpose, *MILSC*=search for comprehension, *MILSS*=search for significance, *MILSP*=search for purpose. Blue lines represent positive partial correlations, red lines represent negative partial correlations

levels in the three search for meaning components. It is also possible to examine each node singularly, for example young people with high average levels of presence of comprehension (MILPC) across the 14 days also showed high levels of search of comprehension (MILSC), but low levels of search for purpose and significance.

Examining the Stability of the Dynamics of the Meaning-Making After 10 Months

The multi-group model showed good fit indices ($\chi^2(7140) = 13029.9$, $p < .001$; CFI = .90; RMSEA = .051 [.049–.052]; AIC = 80374.9; BIC = 81123.9). The three network structures plotted separately for the two waves are represented in Fig. 1. Results indicated that only the contemporaneous matrix was non-invariant across the two waves (Table 2), as constraining the matrices to equality determined an increase of the BIC, suggesting that at least one of the constrained parameters was different across the two waves. To identify the non-invariant parameters, the contemporaneous matrix of the two waves was inspected from the free model. The parameters of contemporaneous matrices in the two waves, together with their confidence intervals can be consulted in supplementary materials. The global path of contemporaneous associations was very similar across the two waves, however, in wave 2 the associations between nodes within the same cluster (presence and

Table 2 Model comparison of multi-group network analysis across the two waves

	X^2	$\Delta\chi^2$	p	RMSEA	AIC	BIC
Model_free	13029.9 (7140)	–	–	.05	80374.9	81123.9
Contemporaneous_constrained	13326.3 (7155)	296.4	<.001	.05	80641.3	81323.4
Temporal_constrained	13085.2 (7176)	89.2	<.001	.05	80392.1	80980.6
Between_constrained	13077.5 (7155)	47.6	<.001	.05	80392.5	81074.6

X^2: Chi-square, $\Delta\chi^2$: Chi-square difference between nested models, *RMSEA*: Root Mean Square Error of Approximation, *AIC*: Akaike Information Criterion, *BIC*: Bayesian Information Criterion

search) were stronger than wave 1, meaning that young people perceived the three components of presence and search for meaning in a much more similar way on the same day while they were living in a more stable contextual condition.

3 Discussion

In the present work, the potentialities of the *psychonetrics* toolbox to investigate the dynamics of change of psychological processes and test their stability over time were examined through the application of the *ml_ts_lvgvar* model, in the basic and the multi-group version. The model converged properly and showed sufficient fit indexes. Considering the complexity of the model and the limited variability of our data (only 14 measurement occasions in each wave) it is noteworthy that the model converged. However, the models used are based on the multiple decomposition of variance, and our sample may not have been large enough, both in terms of sample size and number of assessments, to have good statistical power, especially for the multi-group extension. Results are therefore to be interpreted with caution, and further studies should be conducted on larger samples in order to fully exploit the potential of these statistical models.

A second innovation of the *psychonetrics* toolbox is the extension to test confirmatory hypothesis by conducting multi-group comparisons. With the classical exploratory approach any difference between two groups was usually inspected visually, with *psychonetrics* confirmatory approach, it becomes possible to have an empirical proof of the invariance of parameters across groups. In our illustrative example, we compared each single matrix across the two waves, and we found out that only one matrix (contemporaneous) changed across the measurement occasions. The *psychonetrics* toolbox also offer the possibility to constrain single edges to equality or specific values, thus potentially allowing the identification of even the smallest discrepancies between different groups. In the present study we had to consider the two waves as independent samples to conduct the multi-group comparison, although some participants took part in both waves; in this regard, a statistical reflection on the best way to manage non-independence in repeated measures designs must be opened.

To conclude, the psychometrics toolbox emerged as a promising tool to investigate within and between process dynamics thanks to the multitude of available models dedicated to longitudinal and intensive designs and the possibility of confirmatory testing. *Psychonetrics* is a user-friendly toolbox that can be used by a vast array of researchers in the psychological field, as it doesn't require advanced psychometric and statistical competences to be applied. Indeed, its extension to confirmatory testing directly comes from the SEM framework, from which the use of well-known fit indexes to evaluate the goodness of models directly come. The logic behind the multi-group testing also comes from the SEM framework. From a computational level, *psychonetrics* offers a much easier approach to test multi-group comparisons compared to other recently developed techniques such as permutation methods (e.g., Van Borkulo et al., 2022). The disadvantages of this framework are mainly due to its young age, in fact, computational warnings or errors might occur for which, at the moment, there is not an easy solution. Further simulation studies and applications to real data should be conducted to improve the reliability of the estimates obtained from this kind of data with the aim of promoting its applications to different research designs.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317–332. https://doi.org/10.1007/BF02294359

Arnett, J. J. (2007). Emerging adulthood: What is it, and what is it good for? *Child Development Perspectives, 1*, 68–73. https://doi.org/10.1111/j.1750-8606.2007.00016.x

Borsboom, D., & Cramer, A. O. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology, 9*, 91–121. https://doi.org/10.1017/S0140525X17002266

Borsboom, D., Deserno, M. K., Rhemtulla, M., et al. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers, 1*, 1–18. https://doi.org/10.1038/s43586-021-00055-w

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Epskamp, S. (2020a). Psychonetrics: Structural equation modeling and confirmatory network analysis (R package version 0.10) [computer software]. http://psychonetrics.org/

Epskamp, S. (2020b). Psychometric network models from time-series and panel data. *Psychometrika, 85*, 206–231. https://doi.org/10.1007/s11336-020-09697-

Epskamp, S., Cramer, A. O., Waldorp, L. J., et al. (2012). Qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software, 48*, 1–18. https://doi.org/10.18637/jss.v048.i04

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods, 50*, 195–212. https://doi.org/10.3758/s13428-017-0862-1

Lin, L. C., Huang, P. H., & Weng, L. J. (2017). Selecting path models in SEM: A comparison of model selection criteria. *Structural Equation Modeling: A Multidisciplinary Journal, 24*, 855–869. https://doi.org/10.1080/10705511.2017.1363652

Little, T. D. (2013). *The Oxford handbook of quantitative methods*. Oxford University Press.

Marsman, M., & Rhemtulla, M. (2022). Guest editors' introduction to the special issue "network psychometrics in action": Methodological innovations inspired by empirical problems. *Psychometrika, 87*, 1–11. https://doi.org/10.1007/s11336-022-09861-x

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*, 171–189. https://doi.org/10.1111/j.2044-8317.1985.tb00832.x

Nezlek, J. B. (2017). A practical guide to understanding reliability in studies of within-person variability. *Journal of Research in Personality, 69*, 149–155. https://doi.org/10.1016/j.jrp.2016.06.020

Park, C. L., Riley, K. E., & Snyder, L. B. (2012). Meaning making coping, making sense, and post-traumatic growth following the 9/11 terrorist attacks. *Journal of Positive Psychology, 7*, 198–207. https://doi.org/10.1080/17439760.2012.671347

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

Steger, M. F., Oishi, S., & Kashdan, T. B. (2009). Meaning in life across the life span: Levels and correlates of meaning in life from emerging adulthood to older adulthood. *Journal of Positive Psychology, 4*, 43–52. https://doi.org/10.1080/17439760802303127

Van Borkulo, C. D., van Bork, R., Boschloo, L., et al. (2022). Comparing network structures on three aspects: A permutation test. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000476.

Van Der Maas, H. L., Dolan, C. V., Grasman, R. P., et al. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review, 113*, 842. https://doi.org/10.1037/0033-295X.113.4.842

Zambelli, M., & Tagliabue, S. (2023). The Situational Meaning in Life Evaluation (SMILE): Development and validity evidence of a new integrated measure of meaning in life. *Under Review.*

Index

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8