Psychometrika (2025), **0**, 1–25
doi:10.1017/psy.2025.10046

PSYCHOMETRIC SOCIETY

THEORY AND METHODS

A Novel Method for Detecting Intersectional DIF: Multilevel Random Item Effects Model with Regularized Gaussian Variational Estimation

He Ren[1], Weicong Lyu[2], Chun Wang[1] and Gongjun Xu[3]

[1]College of Education, University of Washington, Seattle, WA, USA; [2]Faculty of Education, University of Macau, Macau, China; [3]Department of Statistics, University of Michigan, Ann Arbor, MI, USA

Corresponding authors: Chun Wang and Gongjun Xu; Email: wang4066@uw.edu, gongjun@umich.edu

(Received 11 March 2025; revised 6 August 2025; accepted 12 August 2025)

Abstract

Differential item functioning (DIF) screening has long been suggested to ensure assessment fairness. Traditional DIF methods typically focus on the main effects of demographic variables on item parameters, overlooking the interactions among multiple identities. Drawing on the intersectionality framework, we define intersectional DIF as deviations in item parameters that arise from the interactions among demographic variables beyond their main effects and propose a novel item response theory (IRT) approach for detecting intersectional DIF. Under our framework, fixed effects are used to account for traditional DIF, while random item effects are introduced to capture intersectional DIF. We further introduce the concept of intersectional impact, which refers to interaction effects on group-level mean ability. Depending on which item parameters are affected and whether intersectional impact is considered, we propose four models, which aim to detect intersectional uniform DIF (UDIF), intersectional UDIF with intersectional impact, intersectional non-uniform DIF (NUDIF), and intersectional NUDIF with intersectional impact, respectively. For efficient model estimation, a regularized Gaussian variational expectation–maximization algorithm is developed. Simulation studies demonstrate that our methods can effectively detect intersectional UDIF, although their detection of intersectional NUDIF is more limited.

Keywords: differential item functioning; intersectional DIF; regularization; variational estimation

1. Introduction

The heavy reliance on assessments in critical social decision-making, such as college admission, personnel selection and placement, and resource allocation, highlights the need for a thorough evaluation of assessment fairness, particularly in light of ongoing concerns about equity. For decades, assessment fairness has been a central focus in psychometrics. American Educational Research Association et al. (2014) further emphasizes the importance of ensuring assessment fairness throughout the test development process, including the standard practice of screening for differential item functioning (DIF).

DIF refers to the phenomenon in which people from different subgroups, usually defined by demographic variables, such as gender, race, or ethnicity, differ in the probability of correctly answering an item after controlling for their ability. Although DIF does not necessarily indicate measurement bias, DIF detection is a critical first step for further investigation. Two types of DIF are often discussed in the literature: uniform DIF (UDIF) and non-uniform DIF (NUDIF). Specifically, UDIF assumes a

consistent difference in item responses between groups across ability levels, whereas NUDIF allows this difference to vary across ability levels. Various DIF detection methods have been developed, including Lord's chi-square test, logistic regression, and regularized DIF (Lord, 1980; Swaminathan & Rogers, 1990; Tutz & Schauberger, 2015; Wang et al., 2023). While these methods differ in many ways, they typically treat DIF as the main effect of each demographic variable.

Recently, two criticisms have emerged concerning the quantitative methodologies used in inequality studies, including those employed in DIF analysis. First, existing methods often overlook intersectionality. In reality, people's multiple identities do not function in isolation but are interlinked to collectively shape the privilege and discrimination. Intersectionality, a theoretical framework rooted in feminist scholarship, highlights this complexity and is increasingly used in fields, such as health, psychology, and education studies (Cole, 2009; Núñez, 2014). In the context of DIF, this framework gives rise to the concept of intersectional DIF, which refers to the DIF that results from the interaction effect of multiple demographic variables. Unlike traditional DIF, which only considers the main effect of demographic variables separately, intersectional DIF captures the potential bias that arises at the intersection of multiple identities. For example, individuals belonging to multiple marginalized groups may experience DIF effects that are not simply the sum of the effect of each grouping variable, but amplified or diminished due to their intersecting social positions. Empirical results from a recent intersectional DIF study suggest that traditional DIF methods that ignore intersectionality may lead to substantial bias (Albano et al., 2024). Second, existing studies often require the specification of a reference group. DIF is usually detected by comparing each focal group to the reference group, while comparisons among focal groups themselves are rarely made. Although mathematically any group can be designated as the reference with no difference, the routine choice of the privileged group may unintentionally reinforce the notion that privileged groups represent the norm, positioning all other groups as deviations (Johfre & Freese, 2021).

In response to these concerns, recent studies have begun to address intersectional DIF (Albano et al., 2024; Belzak, 2023; Muthén & Asparouhov, 2018; Russell & Kaplan, 2021; Russell et al., 2021, 2022). These methods typically model intersectional DIF by either incorporating both demographic variables and their interactions (i.e., product terms) into the measurement model, or by defining a single synthetic categorical variable that encodes all combinations of demographic characteristics. The synthetic group method is mathematically equivalent to modeling all-way interactions. However, both methods treat intersectional DIF as fixed effects, which limits scalability. As more demographic variables are included, the number of intersectional groups and corresponding parameters increases geometrically, while the sample size per group decreases. This leads to challenges for model estimation. For example, the combination of gender (e.g., male, female, and non-binary) and race (e.g., White, Black, or African American, American Indian or Alaska Native, Asian, and Native Hawaiian or Other Pacific Islander) results in 15 intersectional groups, and this number expands rapidly as additional variables are considered.

In contrast to traditional fixed-effect methods for DIF detection, intersectionality can be modeled as random effects within the multilevel modeling framework. This approach is inspired by multilevel analysis of individual heterogeneity and discriminatory accuracy (MAIHDA), an emerging quantitative approach developed in health inequality. MAIHDA treats individuals (level 1) as nested within intersectional strata (level 2), where each stratum represents a unique combination of social identities, that is, a specific level of the synthetic intersectional group variable. MAIHDA models incorporate the main effect for each demographic variable and a stratum-level random effect. Rather than modeling all-way interaction terms explicitly through fixed effects, the random effect captures the total between-strata variance that is not explained by the additive main effects. Compared to traditional fixed-effect methods, this multilevel framework promotes model parsimony and scalability in the presence of many demographic variables and enables the decomposition of covariate effects into additive and interactive components (Evans et al., 2024; Evans et al., 2018; Merlo, 2018). It is worth noting that, in MAIHDA, the same demographic variables that characterize individuals at level 1 also define the level 2 strata. While this seems to introduce collinearity, there is a conceptual distinction. As clarified in the

https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press

MAIHDA literature, unlike in conventional multilevel models where demographic variables are treated as individual-level covariates, these variables are conceptualized as properties of the strata at level 2. This framing is fundamental to the MAIHDA framework and is discussed in detail by Evans et al. (2024).

Similar to the MAIHDA framework, we propose applying random effects to item parameters for detecting intersectional DIF. In this approach, the main effects of demographic variables on item parameters are explicitly modeled to account for traditional DIF, while random effects are introduced to capture additional variations across intersectional groups without requiring the explicit specification of interaction terms. A random item effect with nonzero variance, after controlling for main effects, is interpreted as evidence of intersectional DIF. Using random effects, the proposed model inherits the advantages of MAIHDA, including interpretability, scalability, and parsimony.

Although this multilevel approach is new to DIF detection, it builds on the well-known random item effect framework in psychometrics. Specifically, random-item item response theory (IRT) models allow item parameters to vary across groups following specific distributions. These models have been applied in various measurement invariance contexts, such as longitudinal designs with randomly drawn item samples, international large-scale assessments, and automatic item generation (AIG) or item cloning (De Boeck, 2008; Jong et al., 2007; Lathrop & Cheng, 2017; Muthén & Asparouhov, 2018; Rijmen & Jeon, 2013). However, existing random item effect models cannot be used directly for intersectional DIF detection. First, most existing models define groups using a single demographic variable (e.g., country) and do not involve the decomposition of main and interaction effects. When extended to intersectional groups formed by multiple identities, using only random effects confounds interactions with main effects. In other words, without explicitly modeling main effects, the random effect cannot be directly interpreted as intersectional DIF (Jong et al., 2007; Muthén & Asparouhov, 2018; Rijmen & Jeon, 2013). Second, existing models typically assume random effects on all items, requiring post hoc tests to identify DIF items. Third, model estimation is computationally intensive. Although Rijmen & Jeon (2013) employ variational inference to reduce computational efforts, their algorithm still lacks closed-form solutions and remains computationally demanding.

Our proposed methods address these limitations through three innovations. First, as mentioned above, the proposed models incorporate both main effects of demographic variables and random effects, enabling separation between traditional and intersectional DIF. Second, we impose a log penalty on the random item effects, effectively shrinking the variance to zero for items free from intersectional DIF. Notably, due to this regularization, our methods do not require anchors for random effects. However, anchor items are still needed for main effects, as we assume that both traditional and intersectional DIF can appear on the same item. Since the primary focus of this study is on intersectional DIF, the anchor requirement applies only to main effects and thus plays a limited role. Third, for efficient model estimation, we develop a Gaussian variational expectation–maximization (GVEM) algorithm. Originally introduced to psychometrics for multidimensional IRT (MIRT) estimation, GVEM circumvents the high-dimensional integral in model estimation, achieves a closed-form solution within the EM algorithm, and significantly reduces computational complexity (Cho et al., 2021).

Our methods accommodate both intersectional UDIF and NUDIF detection by applying a unified modeling strategy to different item parameters (i.e., difficulty and discrimination). In addition, we extend the model to capture intersectional impact. Impact refers to differences in the group-level mean abilities. Traditionally, impact is limited to the main effects of demographic variables on group-level ability means. However, as emphasized by intersectionality, interactions among multiple identities could also influence group-level abilities. We define intersectional impact as different group-level mean abilities arising from these interactions. Similar to intersectional DIF, we use random effects on group-level mean abilities (i.e., multilevel latent trait) to capture intersectional impact. Beyond studying intersectionality, our proposed approach is also well suited for nested structures, such as students within different countries, especially in large-scale assessments (Pastor, 2003; Sulis & Toland, 2017).

In summary, the major contributions of this article are fourfold: (1) quantifying the intersectional DIF as random item effects, (2) introducing the concept of intersectional impact, (3) applying a log penalty to detect nonzero item-level variation reflective of intersectional DIF, and (4) applying efficient

https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press

Table 1. Proposed IRT models in this study

	2PL-Ri	2PL-RiM	2PL-Ris	2PL-RisM
Random intercept	Yes	Yes	Yes	Yes
Random slope	No	No	Yes	Yes
Multilevel latent trait	No	Yes	No	Yes
DIF scenario	UDIF	UDIF with impact	UDIF and NUDIF	UDIF and NUDIF with impact

Note: DIF and impact refer to intersectional DIF and intersectional impact, respectively.

variational methods for model estimation. The rest of the article is organized as follows. We first introduce the four random item IRT models proposed in this study, followed by the regularized GVEM algorithm. Then, we present four simulation studies and an empirical study to evaluate the performance of the proposed intersectional DIF detection methods. Finally, we conclude with a discussion of limitations and future directions.

2. Methods

This study aims to detect intersectional DIF, defined as interactions among demographic variables on item parameters. We also consider scenarios both with and without intersectional impact, defined as interactions among demographic variables that affect group-level mean abilities. The two-parameter logistic (2PL) IRT model is used as the foundational model, upon which four extended models are developed. These models incorporate random item intercepts for intersectional UDIF, random item slopes for intersectional NUDIF, and a multilevel ability structure for intersectional impact. Specifically, the four proposed models are 2PL with random item intercept (2PL-Ri), 2PL with random item intercept and with multilevel latent trait (2PL-RiM), 2PL with random item intercept and slope (2PL-Ris), and 2PL with random item intercept and slope and with multilevel latent trait (2PL-RisM). The structure and applicability of these models are summarized in Table 1.

Let y_{ijs} denote the binary response of person i ($i = 1,2,\ldots,N_s$) in group s ($s = 1,2,\ldots,S$) on item j ($j = 1,2,\ldots,J$). For 2PL-RisM, the most flexible model in this study, the item response function of y_{ijs} is

$$\mathbb{P}(y_{ijs} = 1 \mid \theta_{is}, a_{js}, b_{js}) = \frac{1}{1+\exp[-(a_{js}\theta_{is} + b_{js})]}, \tag{1}$$

where the random effects are

$$\begin{aligned} \theta_{is} \mid \alpha_{0s} &\sim \mathcal{N}(\alpha_{0s} + \boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{X}_s, \sigma_\theta^2), \\ \alpha_{0s} &\sim \mathcal{N}(0, \sigma_{\alpha_0}^2), \\ b_{js} &\sim \mathcal{N}(\boldsymbol{\beta}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s, \sigma_{b_j}^2), \\ a_{js} &\sim \mathcal{N}_+(\tilde{\boldsymbol{\gamma}}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s, \tilde{\sigma}_{a_j}^2). \end{aligned} \tag{2}$$

Before interpreting the model, it is necessary to clarify several notations. Let D denote the number of demographic variables of interest. An intersectional group is defined as a unique combination of levels across these D variables. Let S be the total number of such intersectional groups, equal to the product of the number of levels across all variables. That is, $S = \prod_{d=1}^{D} K_d$, where K_d is the number of levels for the d-th variable. For each group $s \in \{1,\ldots,S\}$, let $\boldsymbol{X}_s$ be a P-dimensional dummy coded vector. The total number of dummy variables is $P = \sum_{d=1}^{D}(K_d - 1)$. For example, if there are $D = 2$ variables, race and gender, where race has $K = 5$ categories and gender has $K = 3$ categories, then $S = 5 \times 3 = 15$, and $P = (5-1) + (3-1) = 6$. To accommodate the intercept, let $\tilde{\boldsymbol{X}}_s = [1, \boldsymbol{X}_s^{\mathrm{T}}]^{\mathrm{T}}$ be a $(P+1)$-dimensional vector. Correspondingly, $\boldsymbol{\beta}_j^{\mathrm{T}} = [b_j, \dot{\boldsymbol{\beta}}_j^{\mathrm{T}}]$, where b_j denotes the intercept parameter for item j

https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press

in the reference group, and $\dot{\boldsymbol{\beta}}_j^{\mathrm{T}}$ is the vector of coefficients representing the main effects of demographic variables on the intercept. Note that in the current model, one intersectional group serves as the reference group because demographic variables are dummy coded to capture main effects. However, by using effect coding instead, no single intersectional group is treated as the reference; rather, effects are interpreted relative to the overall mean across all groups. Similarly, $\tilde{\boldsymbol{\gamma}}_j^{\mathrm{T}} = [a_j, \dot{\tilde{\boldsymbol{\gamma}}}_j^{\mathrm{T}}]$, where a_j denotes the slope parameter for item j in the reference group, and $\dot{\tilde{\boldsymbol{\gamma}}}_j^{\mathrm{T}}$ represents the main effects on the slope. Finally, $\mathcal{N}$ and $\mathcal{N}_+$ denote the normal distribution and the truncated normal distribution (left-truncated at zero), respectively. We place bars over parameters associated with a_{js} to indicate that $\tilde{\boldsymbol{\gamma}}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s$ and $\tilde{\sigma}^2_{a_j}$ represent the mean and variance of the untruncated latent variable underlying a_{js}, rather than those of a_{js} itself.

In Equation (1), θ_{is} is the ability of person i in subgroup s, where abilities within each subgroup follow a normal distribution with mean $\alpha_{0s} + \boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{X}_s$. Recall that $\boldsymbol{X}_s$ is the dummy coding vector that corresponds to group s. $\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{X}_s$ represents the main effect of demographics on the group-level mean ability, that is, the traditional impact in DIF literature. In addition, we introduce the random intercept $\alpha_{0s} \sim \mathcal{N}(0, \sigma^2_{\alpha_0})$. As in the MAIHDA literature, we do not explicitly model any high-order interactions among the demographic variables. Instead, the random effect α_{0s} is used to capture these additional deviations, that is, the intersectional impact. Moreover, b_{js} represents the group-specific intercept parameter for item j in subgroup s, following a normal distribution with mean $\boldsymbol{\beta}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s$ and variance $\sigma^2_{b_j}$. The term $\boldsymbol{\beta}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s$ represents the group-specific intercept due to the main effects of demographic variables, corresponding to the traditional UDIF. The variance of the random intercept $\sigma^2_{b_j}$ captures deviations from the main effect across intersectional groups and is intended to reflect intersectional DIF on the intercept. Similarly, the group-specific slope parameter a_{js} follows a truncated normal distribution. Its pre-truncation mean, $\tilde{\boldsymbol{\gamma}}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s$, captures the main effects of demographics on the slope (i.e., traditional NUDIF), while the variance $\tilde{\sigma}^2_{a_j}$ is specifically introduced to capture intersectional DIF on the slope. Accordingly, the model is parameterized so that item j is free of intersectional NUDIF when $\tilde{\sigma}^2_{a_j} = 0$, and further free of intersectional UDIF when $\sigma^2_{b_j} = 0$ as well.

Please note that the 2PL-RisM model shown above combines features of the random item effect model and multilevel IRT model. It treats item parameters similarly to a linear logistic test model with error (LLTM with error), but instead of using a property matrix to explain the difficulty (De Boeck, 2008; Kim & Wilson, 2020), the mean of each item's difficulties is determined by the main effects of demographics to capture traditional DIF, while the variance accounts for additional variations (i.e., intersectional DIF). In addition, the other three models in Table 1 are simplified version of the 2PL-RisM model: the 2PL-Ris model sets $\alpha_{0s} = 0$, the 2PL-RiM model sets $a_{js} = a_j$, and the 2PL-Ri model sets both $\alpha_{0s} = 0$ and $a_{js} = a_j$.

2.1. Model estimation

In this section, we introduce a novel algorithm for model estimation based on variational inference. The key idea of variational approximation is to approximate the intractable marginal likelihood with a computationally feasible lower bound. The lower bound derived in this article follows the local variational methods by Bishop (2006) and Cho et al. (2021). Compared to the variational methods by Rijmen & Jeon (2013), the GVEM method in this article results in closed-form solutions for most parameters. To ensure clarity, we begin with the simplest model in this study, 2PL-Ri. The item response function of y_{ijs} can still be written as in Equation (1), with random effects

$$\begin{aligned} \theta_{is} &\sim \mathcal{N}(\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{X}_s, \sigma^2_\theta), \\ b_{js} &\sim \mathcal{N}(\boldsymbol{\beta}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s, \sigma^2_{b_j}). \end{aligned} \tag{3}$$

Let $\boldsymbol{Z} = \bigcup_{s=1}^{S}\bigcup_{i=1}^{N_s}\bigcup_{j=1}^{J}\{\theta_{is}, b_{js}\}$ be the set of all latent variables, including both latent traits and random item effects, in the 2PL-Ri model. The joint likelihood of responses $\boldsymbol{Y} = \bigcup_{s=1}^{S}\bigcup_{i=1}^{N_s}\bigcup_{j=1}^{J}\{y_{ijs}\}$ and latent

https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press

variables $\boldsymbol{Z}$ is

$$\begin{aligned} p(\boldsymbol{Y},\boldsymbol{Z}) &= \mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{Z}) p(\boldsymbol{Z}) \\ &= \prod_{s=1}^{S} \left\{ \left[\prod_{i=1}^{N_s} \prod_{j=1}^{J} \mathbb{P}(Y_{ijs} = y_{ijs} \mid \theta_{is}, b_{js}) \right] \left[\prod_{j=1}^{J} p_{b_j}(b_{js}) \right] \left[\prod_{i=1}^{N_s} p_{\theta_{is}}(\theta_{is}) \right] \right\}, \end{aligned} \tag{4}$$

where $\mathbb{P}(Y_{ijs} = y_{ijs} \mid \theta_{is}, b_{js}) = \mathbb{P}(Y_{ijs} = 1 \mid \theta_{is}, b_{js})^{y_{ijs}} \mathbb{P}(Y_{ijs} = 0 \mid \theta_{is}, b_{js})^{1-y_{ijs}}$. With any probability density function $q(\boldsymbol{Z})$ for $\boldsymbol{Z}$, the log marginal likelihood of $\boldsymbol{Y}$ can be written as

$$\begin{aligned} \ell(\boldsymbol{Y}) &= \log \mathbb{P}(\boldsymbol{Y}) \\ &= \int [\log \mathbb{P}(\boldsymbol{Y})] q(\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z} \\ &= \int \left[\log \frac{p(\boldsymbol{Y},\boldsymbol{Z})}{p(\boldsymbol{Z} \mid \boldsymbol{Y})} \right] q(\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z} \\ &= \int \left[\log \frac{p(\boldsymbol{Y},\boldsymbol{Z})}{q(\boldsymbol{Z})} \right] q(\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z} + \int \left[\log \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z} \mid \boldsymbol{Y})} \right] q(\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z} \\ &\geq \int \left[\log \frac{p(\boldsymbol{Y},\boldsymbol{Z})}{q(\boldsymbol{Z})} \right] q(\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z} \\ &\equiv \mathrm{ELBO} \\ &= \int [\log p(\boldsymbol{Y},\boldsymbol{Z})] q(\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z} - \text{constant}, \end{aligned} \tag{5}$$

where ELBO refers to the evidence lower bound, and the difference $\ell(\boldsymbol{Y}) - \mathrm{ELBO}$ corresponds to the Kullback–Leibler (KL) divergence (Kullback & Leibler, 1951) from $q(\boldsymbol{Z})$ to $p(\boldsymbol{Z} \mid \boldsymbol{Y})$, given by

$$\mathrm{KL}[q(\boldsymbol{Z}) \parallel p(\boldsymbol{Z} \mid \boldsymbol{Y})] = \int \left[\log \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z} \mid \boldsymbol{Y})} \right] q(\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z} \geq 0.$$

Note that the constant in Equation (5), $\int [\log q(\boldsymbol{Z})] q(\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z}$, depends only on q and can therefore be omitted from the optimization. Optimizing $\ell(\boldsymbol{Y})$ is thus reduced to maximizing $\int [\log p(\boldsymbol{Y},\boldsymbol{Z})] q(\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z}$. The EM algorithm achieves this by setting $q(\boldsymbol{Z})$ such that $\mathrm{KL}[q(\boldsymbol{Z}) \parallel p(\boldsymbol{Z} \mid \boldsymbol{Y})] = 0$. In the E-step, it computes the expectation of the log-likelihood (i.e., $\int [\log p(\boldsymbol{Y},\boldsymbol{Z})] q(\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z}$). In the M-step, this expectation is maximized with respect to model parameters. However, the regular EM algorithm requires that the expectation is computationally feasible, which hardly holds in the random item effect models. In the 2PL-Ri model, for example, the expectation in Equation (5) involves a high-dimensional integral with respect to $\boldsymbol{Z}$, a latent variable of dimension $SJ + N$, where $N = \sum_{s=1}^{S} N_s$ is the total sample size. We address this challenge by applying variational inference for estimation.

In the context of 2PL-Ri, given Equations (1) and (3)–(4), we have

$$\begin{aligned} &\int [\log p(\boldsymbol{Y},\boldsymbol{Z})] q(\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z} \\ &= \int [\log \mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{Z}) + \log p(\boldsymbol{Z})] q(\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z} \\ &= \int \left[\sum_{s=1}^{S} \sum_{i=1}^{N_s} \sum_{j=1}^{J} \left\{ y_{ijs} \log \frac{1}{1 + \exp[-(a_j \theta_{is} + b_{js})]} + (1 - y_{ijs}) \log \frac{1}{1 + \exp(a_j \theta_{is} + b_{js})} \right\} \right. \\ &\qquad \left. + \sum_{s=1}^{S} \left\{ \sum_{i=1}^{N_s} \log p_{\theta_{is}}(\theta_{is}) + \sum_{j=1}^{J} \log p_{b_{js}}(b_{js}) \right\} \right] q(\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z} \end{aligned} \tag{6}$$

https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press

$$= \int \left[\sum_{s=1}^{S} \sum_{i=1}^{N_s} \sum_{j=1}^{J} \left\{ y_{ijs} \log \frac{1}{1+\exp[-(a_j\theta_{is}+b_{js})]} + (1-y_{ijs}) \log \frac{1}{1+\exp(a_j\theta_{is}+b_{js})} \right\} \right.$$
$$- \frac{1}{2} \sum_{s=1}^{S} \left\{ (N_s+J)\log 2\pi + \sum_{i=1}^{N_s} \left[\log \sigma_\theta^2 + \frac{(\theta_{is} - \boldsymbol{\alpha}_1^{\mathrm{T}} \boldsymbol{X}_s)^2}{\sigma_\theta^2} \right] \right.$$
$$\left. \left. + \sum_{j=1}^{J} \left[\log \sigma_{b_j}^2 + \frac{(b_{js} - \boldsymbol{\beta}_j^{\mathrm{T}} \tilde{\boldsymbol{X}}_s)^2}{\sigma_{b_j}^2} \right] \right\} \right] q(\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z}.$$

The difficulty in computing the marginal log-likelihood in Equation (6) primarily arises from the sigmoid function, which prevents closed-form integration. We adopted a local variational method (Bishop, 2006) to approximate the sigmoid function with a computationally feasible lower bound. As demonstrated in Cho et al. (2021), a sigmoid function can be expressed as

$$\begin{aligned} f(x) = \frac{\exp(x)}{1+\exp(x)} &= \max_{\xi} \frac{\exp(\xi)}{1+\exp(\xi)} \exp\left[\frac{(x-\xi)}{2} - \eta(\xi)(x^2-\xi^2) \right] \\ &\geq \frac{\exp(\xi)}{1+\exp(\xi)} \exp\left[\frac{(x-\xi)}{2} - \eta(\xi)(x^2-\xi^2) \right], \end{aligned} \tag{7}$$

where $\eta(\xi) = (2\xi)^{-1}\{1/[1+\exp(-\xi)] - 1/2\}$, and ξ is the variational parameter used to approximate the sigmoid function, which is updated iteratively in the EM algorithm. Applying Equation (7) to Equation (6), we obtain

$$\begin{aligned} &\int [\log p(\boldsymbol{Y},\boldsymbol{Z})]\, q(\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z} \\ &\geq \int \left[\sum_{s=1}^{S} \sum_{i=1}^{N_s} \sum_{j=1}^{J} \left\{ \log \frac{1}{1+\mathrm{e}^{-\xi_{ijs}}} + \left(y_{ijs} - \frac{1}{2} \right)(a_j\theta_{is} - b_{js}) - \frac{1}{2}\xi_{ijs} - \eta(\xi_{ijs})\left[(a_j\theta_{is} - b_{js})^2 - \xi_{ijs}^2 \right] \right\} \right. \\ &\quad - \frac{1}{2} \sum_{s=1}^{S} \left\{ (N_s+J)\log 2\pi + \sum_{i=1}^{N_s} \left[\log \sigma_\theta^2 + \frac{(\theta_{is} - \boldsymbol{\alpha}_1^{\mathrm{T}} \boldsymbol{X}_s)^2}{\sigma_\theta^2} \right] \right. \\ &\quad \left. \left. + \sum_{j=1}^{J} \left[\log \sigma_{b_j}^2 + \frac{(b_{js} - \boldsymbol{\beta}_j^{\mathrm{T}} \tilde{\boldsymbol{X}}_s)^2}{\sigma_{b_j}^2} \right] \right\} \right] q(\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z} \\ &\equiv \int \mathcal{B}(\boldsymbol{Y},\boldsymbol{Z}) q(\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z} \\ &\equiv Q(\boldsymbol{Y}), \end{aligned} \tag{8}$$

where $\mathcal{B}(\boldsymbol{Y},\boldsymbol{Z})$ denotes a lower bound of $\log p(\boldsymbol{Y},\boldsymbol{Z})$ under the local variational approximation.

Next, we need to determine a variational density $q(\boldsymbol{Z})$ that closely approximates the true posterior $p(\boldsymbol{Z} \mid \boldsymbol{Y})$, such that $\mathrm{KL}[q(\boldsymbol{Z}) \parallel p(\boldsymbol{Z} \mid \boldsymbol{Y})]$ is minimized. This ensures that the ELBO provides a tight approximation to the marginal log-likelihood $\ell(\boldsymbol{Y})$, as shown in Equation (5). Under the mean-field variational assumption (Bishop, 2006), we approximate the posterior distribution of the latent variables using a product of independent factors, each corresponding to a separate latent variable, i.e.,

$$q(\boldsymbol{Z}) = \prod_{s=1}^{S} \left[\prod_{i=1}^{N_s} q_{\theta_{is}}(\theta_{is}) \prod_{j=1}^{J} q_{b_{js}}(b_{js}) \right].$$

Note that the latent variables in $\boldsymbol{Z}$ need not be truly independent, as the goal is to approximate its true posterior distribution while simplifying the computation. Then, for any latent variable $z_\ell \in \boldsymbol{Z}$, its optimal variational distribution q_{z_ℓ} takes the form

$$q_{z_\ell}(z_\ell) \propto \exp \mathbb{E}_{\boldsymbol{Z} \setminus \{z_\ell\}} [\log p(\boldsymbol{Y},\boldsymbol{Z})],$$

https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press

where $\mathbb{E}_{\boldsymbol{Z}\setminus\{z_\ell\}}$ refers to the expectation over all latent variables in $\boldsymbol{Z}$ other than z_ℓ (Bishop, 2006; Blei et al., 2017). With the lower bound $\mathcal{B}(\boldsymbol{Y},\boldsymbol{Z})$, we update the variational distribution as

$$q_{z_\ell}(z_\ell) \propto \exp \mathbb{E}_{\boldsymbol{Z}\setminus\{z_\ell\}}[\mathcal{B}(\boldsymbol{Y},\boldsymbol{Z})].$$

Thus, the optimal $q_{\theta_{is}}(\theta_{is})$ and $q_{b_{js}}(b_{js})$ that maximize the ELBO (i.e., minimize the KL divergence) are given by

$$\begin{aligned}
q_{\theta_{is}}(\theta_{is}) &\propto \exp \mathbb{E}_{\boldsymbol{Z}\setminus\{\theta_{is}\}}[\mathcal{B}(\boldsymbol{Y},\boldsymbol{Z})] \\
&\propto \exp \mathbb{E}_{\boldsymbol{Z}\setminus\{\theta_{is}\}} \left\{ \sum_{j=1}^{J} \left[\left(y_{ijs} - \frac{1}{2} \right) a_j \theta_{is} - \eta(\xi_{ijs}) (a_j \theta_{is} - b_{js})^2 \right] - \frac{(\theta_{is} - \boldsymbol{\alpha}_1^{\mathrm{T}} \boldsymbol{X}_s)^2}{2\sigma_\theta^2} \right\} \\
&\propto \exp \left\{ \sum_{j=1}^{J} \left[\left(y_{ijs} - \frac{1}{2} \right) a_j \theta_{is} - \eta(\xi_{ijs}) (a_j^2 \theta_{is}^2 - 2 a_j \theta_{is} \mu_{b_{js}} + \mu_{b_{js}}^2 + \sigma_{b_{js}}^2) \right] - \frac{(\theta_{is} - \boldsymbol{\alpha}_1^{\mathrm{T}} \boldsymbol{X}_s)^2}{2\sigma_\theta^2} \right\} \\
&\propto \exp \left\{ \sum_{j=1}^{J} a_j \left[\left(y_{ijs} - \frac{1}{2} \right) + 2\eta(\xi_{ijs}) \mu_{b_{js}} \right] \theta_{is} - \sum_{j=1}^{J} \left[\eta(\xi_{ijs}) a_j^2 \right] \theta_{is}^2 - \frac{\theta_{is}^2 - 2\boldsymbol{\alpha}_1^{\mathrm{T}} \boldsymbol{X}_s \theta_{is}}{2\sigma_\theta^2} \right\} \\
&\propto \exp \left\{ \left[\sum_{j=1}^{J} a_j \left\{ \left(y_{ijs} - \frac{1}{2} \right) + 2\eta(\xi_{ijs}) \mu_{b_{js}} \right\} + \frac{\boldsymbol{\alpha}_1^{\mathrm{T}} \boldsymbol{X}_s}{\sigma_\theta^2} \right] \theta_{is} - \left[\sum_{j=1}^{J} \eta(\xi_{ijs}) a_j^2 + \frac{1}{2\sigma_\theta^2} \right] \theta_{is}^2 \right\}
\end{aligned} \tag{9}$$

and

$$\begin{aligned}
q_{b_{js}}(b_{js}) &\propto \exp \mathbb{E}_{\boldsymbol{Z}\setminus\{b_{js}\}}[\mathcal{B}(\boldsymbol{Y},\boldsymbol{Z})] \\
&\propto \exp \mathbb{E}_{\boldsymbol{Z}\setminus\{b_{js}\}} \left\{ -\sum_{i=1}^{N_s} \left[\left(y_{ijs} - \frac{1}{2} \right) b_{js} + \eta(\xi_{ijs}) (b_{js} - a_j \theta_{is})^2 \right] - \frac{(b_{js} - \boldsymbol{\beta}_j^{\mathrm{T}} \tilde{\boldsymbol{X}}_s)^2}{2\sigma_{b_j}^2} \right\} \\
&\propto \exp \left\{ -\sum_{i=1}^{N_s} \left[\left(y_{ijs} - \frac{1}{2} \right) b_{js} + \eta(\xi_{ijs}) \left\{ b_{js}^2 - 2 b_{js} a_j \mu_{\theta_{is}} + a_j^2 (\mu_{\theta_{is}}^2 + \sigma_{\theta_{is}}^2) \right\} \right] - \frac{(b_{js} - \boldsymbol{\beta}_j^{\mathrm{T}} \tilde{\boldsymbol{X}}_s)^2}{2\sigma_{b_j}^2} \right\} \\
&\propto \exp \left\{ \sum_{i=1}^{N_s} \left[2\eta(\xi_{ijs}) a_j \mu_{\theta_{is}} - \left(y_{ijs} - \frac{1}{2} \right) \right] b_{js} - \sum_{i=1}^{N_s} \eta(\xi_{ijs}) b_{js}^2 - \frac{b_{js}^2 - 2\boldsymbol{\beta}_j^{\mathrm{T}} \tilde{\boldsymbol{X}}_s b_{js}}{2\sigma_{b_j}^2} \right\} \\
&\propto \exp \left\{ \left[\sum_{i=1}^{N_s} \left\{ 2\eta(\xi_{ijs}) a_j \mu_{\theta_{is}} - \left(y_{ijs} - \frac{1}{2} \right) \right\} + \frac{\boldsymbol{\beta}_j^{\mathrm{T}} \tilde{\boldsymbol{X}}_s}{\sigma_{b_j}^2} \right] b_{js} - \left[\sum_{i=1}^{N_s} \eta(\xi_{ijs}) + \frac{1}{2\sigma_{b_j}^2} \right] b_{js}^2 \right\},
\end{aligned} \tag{10}$$

respectively. As shown in Equation (9), the variational density of θ_{is} is an exponential family with sufficient statistics θ_{is} and θ_{is}^2, and thus $q_{\theta_{is}}(\theta_{is}) = \mathcal{N}(\mu_{\theta_{is}}, \sigma_{\theta_{is}}^2)$, where

$$\begin{aligned}
\sigma_{\theta_{is}}^2 &= \frac{\sigma_\theta^2}{1 + 2\sigma_\theta^2 \sum_{j=1}^{J} \eta(\xi_{ijs}) a_j^2} \\
\mu_{\theta_{is}} &= \frac{\boldsymbol{\alpha}_1^{\mathrm{T}} \boldsymbol{X}_s + \sigma_\theta^2 \sum_{j=1}^{J} a_j \left[y_{ijs} - \frac{1}{2} + 2\eta(\xi_{ijs}) \mu_{b_{js}} \right]}{1 + 2\sigma_\theta^2 \sum_{j=1}^{J} \eta(\xi_{ijs}) a_j^2}.
\end{aligned} \tag{11}$$

Similarly, the variational density of b_{is} shown in Equation (10) also follows a normal distribution, that is, $q_{b_{is}}(b_{is}) = \mathcal{N}(\mu_{b_{js}}, \sigma_{b_{js}}^2)$, where

$$\begin{aligned}
\sigma_{b_{js}}^2 &= \frac{\sigma_{b_j}^2}{1 + 2\sigma_{b_j}^2 \sum_{i=1}^{N_s} \eta(\xi_{ijs})} \\
\mu_{b_{js}} &= \frac{\boldsymbol{\beta}_j^{\mathrm{T}} \tilde{\boldsymbol{X}}_s - \sigma_{b_j}^2 \sum_{i=1}^{N_s} \left[y_{ijs} - \frac{1}{2} - 2\eta(\xi_{ijs}) a_j \mu_{\theta_{is}} \right]}{1 + 2\sigma_{b_j}^2 \sum_{i=1}^{N_s} \eta(\xi_{ijs})}.
\end{aligned} \tag{12}$$

https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press

Given the optimal variational distributions derived above and the mean-field variational assumption, we compute the expectation over all latent variables with respect to the variational distribution $q(\boldsymbol{Z})$ in Equation (8), yielding

$$\begin{aligned} Q(\boldsymbol{Y}) = &\sum_{s=1}^{S}\sum_{i=1}^{N_s}\sum_{j=1}^{J}\left\{\log\frac{1}{1+\mathrm{e}^{-\xi_{ijs}}}+\left(y_{ijs}-\frac{1}{2}\right)\left(a_j\mu_{\theta_{is}}-\mu_{b_{js}}\right)-\frac{1}{2}\xi_{ijs}\right. \\ &\left.-\eta(\xi_{ijs})\left[a_j^2\left(\mu_{\theta_{is}}^2+\sigma_{\theta_{is}}^2\right)-2a_j\mu_{\theta_{is}}\mu_{b_{js}}+\mu_{b_{js}}^2+\sigma_{b_{js}}^2-\xi_{ijs}^2\right]\right\} \\ &-\frac{1}{2}\sum_{s=1}^{S}\left\{(N_s+J)\log 2\pi+\sum_{i=1}^{N_s}\left[\log\sigma_\theta^2+\frac{\sigma_{\theta_{is}}^2+\left(\mu_{\theta_{is}}-\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{X}_s\right)^2}{\sigma_\theta^2}\right]\right. \\ &\left.+\sum_{j=1}^{J}\left[\log\sigma_{b_j}^2+\frac{\sigma_{b_{js}}^2+\left(\mu_{b_{js}}-\boldsymbol{\beta}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s\right)^2}{\sigma_{b_j}^2}\right]\right\}. \end{aligned} \tag{13}$$

In addition, a log penalty is imposed on $\sigma_{b_j}^2$ in Equation (13) to encourage sparsity in item random effects for intersectional DIF detection. The log penalty has been employed for identifying permissible attribute patterns in cognitive diagnostic models (Gu & Xu, 2019; Ma et al., 2023; Wang, 2024). Note that $\log\sigma_{b_j}^2$ is already included in $Q(\boldsymbol{Y})$. On the one hand, incorporating the log penalty preserves closed-form solutions in the M-step, thereby ensuring computational efficiency. On the other hand, as shown by Ma et al. (2023), the log penalty has a Bayesian interpretation: it corresponds to placing a Dirichlet prior with parameter $1-\lambda$ on the variances. When $1-\lambda<0$, the prior becomes an improper Dirichlet distribution, which promotes the selection of significant variances more aggressively than traditional proper Dirichlet priors. Overall, a regularized GVEM algorithm is proposed, where the objective function to be maximized is given by

$$Q'(\boldsymbol{Y}) = Q(\boldsymbol{Y}) - \lambda\sum_{j=1}^{J}\log\sigma_{b_j}^2,$$

where $\lambda>0$ is a tuning parameter and larger values of λ result in greater sparsity in $\sigma_{b_j}^2$.

In each EM iteration, variational densities in Equations (11) and (12) are updated in the E-step. In the M-step, $Q'(\boldsymbol{Y})$ is maximized to update all model parameters. This is achieved by setting the derivative of the objective function with respect to each model parameter to be zero. We will show that all parameters of the 2PL-Ri model can be updated with closed-form solutions, leading to a computationally efficient algorithm. We fix σ_θ to 1 for model identification, and the update rules for all other model parameters are presented below:

$$\begin{aligned} \xi_{ijs}^2 &= \mu_{b_{js}}^2+\sigma_{b_{js}}^2-2a_j\mu_{\theta_{is}}\mu_{b_{js}}+a_j^2\left(\mu_{\theta_{is}}^2+\sigma_{\theta_{is}}^2\right), \\ a_j &= \frac{\sum_{s=1}^{S}\sum_{i=1}^{N_s}\left[\left(y_{ijs}-\frac{1}{2}\right)\mu_{\theta_{is}}+2\eta(\xi_{ijs})\mu_{\theta_{is}}\mu_{b_{js}}\right]}{2\sum_{s=1}^{S}\sum_{i=1}^{N_s}\eta(\xi_{ijs})\left(\mu_{\theta_{is}}^2+\sigma_{\theta_{is}}^2\right)}, \\ \boldsymbol{\alpha}_1 &= \left(\sum_{s=1}^{S}\sum_{i=1}^{N_s}\boldsymbol{X}_s\boldsymbol{X}_s^{\mathrm{T}}\right)^{-1}\left(\sum_{s=1}^{S}\sum_{i=1}^{N_s}\mu_{\theta_{is}}\boldsymbol{X}_s\right), \\ \boldsymbol{\beta}_j &= \left(\sum_{s=1}^{S}\tilde{\boldsymbol{X}}_s\tilde{\boldsymbol{X}}_s^{\mathrm{T}}\right)^{-1}\left(\sum_{s=1}^{S}\mu_{b_{js}}\tilde{\boldsymbol{X}}_s\right), \\ \sigma_{b_j}^2 &= \frac{1}{S+2\lambda}\sum_{s=1}^{S}\left[\sigma_{b_{js}}^2+\left(\boldsymbol{\beta}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s-\mu_{b_{js}}\right)^2\right]. \end{aligned} \tag{14}$$

https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press

Following the derivation shown above, similar variational lower bounds can be derived for the other three proposed models. The detailed derivation for the most complex model in this study, 2PL-RisM, is provided in the Appendix.

Lastly, we employ the generalized information criterion (GIC) to select an appropriate value of λ, as it has been shown to have desirable theoretical properties (Cho et al., 2024; Fan & Tang, 2013). Specifically, GIC takes the form of

$$\text{GIC} = -2\ell(\boldsymbol{Y}) + k \times c_N, \tag{15}$$

where k is the number of DIF parameters and $c_N = c \log N \log\log N$, with c being a constant that controls the degree of model sparsity. When $c_N = \log N$, GIC reduces to the Bayesian information criterion (BIC). Since $\ell(\boldsymbol{Y})$ is computationally intractable due to high-dimensional integration, we instead use $Q(\boldsymbol{Y})$ as a surrogate in Equation (15) to compute the GIC.

The regularized GVEM algorithm for DIF detection in the 2PL-Ri model is summarized in Algorithm 1. Two remarks are worth noting.

Remark 1. The log penalty term $\log x$ might lead to numerical instability when x approaches zero. Although $\log x$ does not appear in the iterations of the GVEM algorithm as shown in Equations (11), (12), and (14), it is required for computing the GIC. Therefore, we replace $\sigma^2_{b_j}$ with $\max\{0.1, \sigma^2_{b_j}\}$ in the GIC calculation whenever $\sigma_{b_j} \neq 0$.

Remark 2. We do not penalize main DIF effects (i.e., $\dot{\boldsymbol{\beta}}_j$ and/or $\dot{\boldsymbol{\gamma}}_j$) because the primary goal of this study is to detect intersectional DIF, which is defined through nonzero variance terms. As a result, anchor items must be prespecified, in contrast to approaches in the literature that penalize main effects directly (see, e.g., Wang et al., 2023). If an additional penalty term, such as the lasso, were imposed on these main effects, anchor items would no longer be required. In this study, we use four anchor items, corresponding to 20% of the total test length.

Algorithm 1 Regularized GVEM algorithm for DIF detection in the 2PL-Ri model.

$\text{GIC}_{\text{best}} \leftarrow +\infty$
for each value of λ **do**
  Initialize all model parameters: $\xi^2_{ijs}, \boldsymbol{\alpha}_1, a_j, \boldsymbol{\beta}_j, \sigma^2_{b_j}, \sigma^2_\theta$
  while not converged **do** ▷ *Convergence is achieved when the maximum change in parameter estimates across successive EM iterations is less than* $\varepsilon = 0.001$
    E-step: Given current model parameters, update the means and variances of the variational distributions using Equations (11) and (12)
    M-step: Given current variational distributions, update model parameters using Equation (14)
  $k_\lambda \leftarrow 0$ ▷ *Count the number of nonzero variance parameters*
  for $j \leftarrow 1$ **to** J **do**
    if $\sigma^2_{b_j} < \rho_b$ **then** ▷ *Threshold set to* $\rho_b = 0.001$
      $\sigma^2_{b_j} \leftarrow 0$
    else
      $k_\lambda \leftarrow k_\lambda + 1$
  Rerun the EM algorithm without penalty (i.e., set $\lambda = 0$), allowing only items with $\sigma_{b_j} \neq 0$ to retain random item effects, to obtain the final estimates
  $\text{GIC}_\lambda \leftarrow -2Q(\boldsymbol{Y}) + kc \log N \log\log N$
  if $\text{GIC}_{\text{best}} > \text{GIC}_\lambda$ **then**
    $\text{GIC}_{\text{best}} \leftarrow \text{GIC}_\lambda$
    Store current parameter estimates as optimal

https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press

3. Simulation studies

Four simulation studies are conducted to evaluate the performance of the proposed regularized GVEM algorithm in detecting intersectional DIF. Studies I–IV corresponded to the four models, 2PL-Ri, 2PL-RiM, 2PL-Ris, and 2PL-RisM, respectively, each targeting at a different DIF scenario, as detailed in the Methods section. In all studies, the number of items is fixed to $J = 20$. Following Huang et al. (2024), the slope parameters a_j ($j = 1,2,\ldots,J$) are drawn from Lognormal$(0,0.25^2)$, and intercept b_j ($j = 1,2,\ldots,J$) are drawn from Uniform$[-2,2]$. The true item parameters are given in Table 2.

Each simulation study systematically manipulates four common factors. First, the number of intersectional groups is set to either 10 or 40. For the 10-group conditions (i.e., $S = 10$), groups are defined by two demographic variables, one binary (e.g., sex) and one five-category variable (e.g., occupational status), resulting in $P = 5$ for the dummy-coded variables. For the 40-group conditions, groups are defined by four demographic variables, three binary (e.g., sex, immigrant background, and dichotomous education level) and one five-category variable (e.g., occupational status), resulting in $P = 7$ for the dummy-coded variables. This setup aligns with an empirical study on intersectionality (Keller et al., 2023). Second, the sample size per group is set to either 50 or 100. Third, the proportion of items with intersectional DIF is set at 20% (Items 1–4) or 60% (Items 1–12). Intersectional DIF is introduced by nonzero random item effects. For items with intersectional UDIF, half are assigned $\sigma^2_{b_j} = 0.54$ and the other half $\sigma^2_{b_j} = 1$. For items with intersectional NUDIF, half are assigned $\tilde{\sigma}^2_{a_j} = 0.33$ and the other half $\tilde{\sigma}^2_{a_j} = 1$. These magnitudes are derived from a pilot study using PISA data, where the variances for intersectional DIF ranged from 0.31 to 0.98 for intercepts (centered at 0.54) and from 0.22 to 0.46 for slopes (centered at 0.33). All items with intersectional NUDIF include both random intercept and random slope, reflecting real-world scenarios where NUDIF often coexists with UDIF (Wang et al., 2023). Fourth, traditional impact, defined as mean ability differences due to the main effects by demographic variables, is either absent or present. When present, the traditional impact is set at $\boldsymbol{\alpha}_1 = 0.1$, yielding ability mean differences ranging from 0.1 to 0.2 for the 10-group conditions and up to 0.4 for the 40-group conditions.

Beyond the four common factors, Studies II and IV also consider intersectional impact. This is introduced by the variance of the group-level random intercept on ability ($\sigma^2_{\alpha_0}$), set to either 0 (absence) or 0.5 (presence). These values correspond to intra-class correlation (ICC) values of approximately 0 and 0.1, aligning with the ICC in empirical intersectional educational assessment literature (Keller et al., 2023). Overall, Studies I and III included 16 experimental conditions each, while Studies II and IV included 32 conditions each. A summary of all manipulated factors is provided in Table 3.

We note again that this study focuses on detecting intersectional DIF, rather than traditional DIF. To avoid confounding due to traditional DIF and to demonstrate the models' ability to disentangle traditional and intersectional DIF, all items are designed to exhibit traditional DIF, which is introduced through fixed main effects of demographic variables. More specifically, we set $\boldsymbol{\beta}_j^{\mathrm{T}} = [b_j, 0.2 \times 1_P^{\mathrm{T}}]$ and $\tilde{\boldsymbol{\gamma}}_j^{\mathrm{T}} = [\tilde{a}_j, 0.1 \times 1_P^{\mathrm{T}}]$ in Equation (2). Under the 10-group conditions, this setup results in traditional DIF magnitudes ranging from 0.2 to 0.4 for intercepts and 0.1 to 0.2 for slopes across groups. For the

Table 2. True fixed item parameters for the simulation studies

Item	1	2	3	4	5	6	7	8	9	10
a_j	0.691	1.483	0.787	0.795	0.607	0.934	0.924	0.855	0.974	1.113
b_j	0.354	0.122	1.911	−1.209	1.377	−1.620	−0.475	−1.816	−1.390	1.099
Item	11	12	13	14	15	16	17	18	19	20
a_j	0.823	0.724	0.823	1.003	0.963	0.839	1.346	1.089	1.135	0.929
b_j	−0.422	−0.554	−0.316	−0.712	0.209	1.885	0.232	0.297	0.565	1.296

https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press

Table 3. Illustration of simulation designs

	Simulation I (2PL-Ri)	Simulation II (2PL-RiM)	Simulation III (2PL-Ris)	Simulation IV (2PL-RisM)
Number of groups (S)	10, 40	10, 40	10, 40	10, 40
Sample size per group (N_s)	50, 100	50, 100	50, 100	50, 100
Proportion of DIF items	20%, 60%	20%, 60%	20%, 60%	20%, 60%
Traditional impact ($\boldsymbol{\alpha}_1$)	0.1, 0	0.1, 0	0.1, 0	0.1, 0
Intersectional impact ($\sigma^2_{\alpha_0}$)	—	0.5, 0	—	0.5, 0

40-group conditions, these ranges increase to 0.2 to 0.8 for intercepts and 0.1 to 0.4 for slopes, following the design by Belzak & Bauer (2020). To ensure model identification in the presence of traditional DIF across all items, 20% of items (Items 17–20) are designated as anchors with main effects fixed at zero (i.e., $\dot{\boldsymbol{\beta}}_j = 0$ and $\dot{\boldsymbol{\gamma}} = 0$). However, their random effects are still freely estimated, meaning that they are not anchored with respect to intersectional DIF.

The flagging procedure for intersectional DIF has been shown in Algorithm 1. Each condition is replicated 50 times, and the performance is measured by false positive (FP) and true positive (TP) rates. Specifically, the FP rate refers to the proportion of items free from intersectional DIF mistakenly flagged as DIF items, while the TP rate refers to the proportion of items with intersectional DIF that are correctly detected. We consider 50 replications sufficient since the TP and FP rates are averaged across all the DIF-free and DIF-related item parameters, rather than being evaluated for a single parameter in each replication.

3.1. Simulation I: UDIF detection

We evaluate 2PL-Ri in this simulation, where slope parameters are fixed across groups. Figure 1 shows the TP and FP rates of Simulation I across 50 replications. Under most conditions, except when $S = 10$ and $N_s = 50$, the new method performs well. Overall, the 2PL-Ri model performs better with more groups and larger sample sizes per group. As intersectional DIF is modeled by random effects, such results are not surprising but consistent with the findings from the multilevel modeling literature (Adam et al., 2021; Maas & Hox, 2005; Moineddin et al., 2007). In addition, the proportion of DIF items and the presence or absence of intersectional impact have minimal influence on the results.

3.2. Simulation II: UDIF detection with intersectional impact

We study 2PL-RiM in Simulation II, where intersectional impact is considered. As shown in Figure 2, the new method follows a pattern similar to Simulation I. That is, the proposed method performs well under most conditions except when $S = 10$ and $N_s = 50$, and unsurprisingly, it performs better with more groups and larger sample sizes per group. In addition, the proportion of DIF items has a small effect on performance, with a lower proportion yielding slightly better results. Similarly, the presence or absence of traditional and intersectional impact has minimal influence on the results.

3.3. Simulation III: UDIF and NUDIF detection

2PL-Ris is evaluated with both UDIF and NUDIF incorporated. Figure 3a and 3b summarizes the DIF detection results on intercept and slope parameters, respectively. With intersectional NUDIF, the proposed method maintains desirable performance on intercepts with $S = 40$ while resulting in worse performance with the smaller group number $S = 10$. The DIF detection results for slope parameters are generally unsatisfactory. Relatively better performance was observed under conditions involving a large number of groups and either the absence of traditional impact or the combination of the

https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press

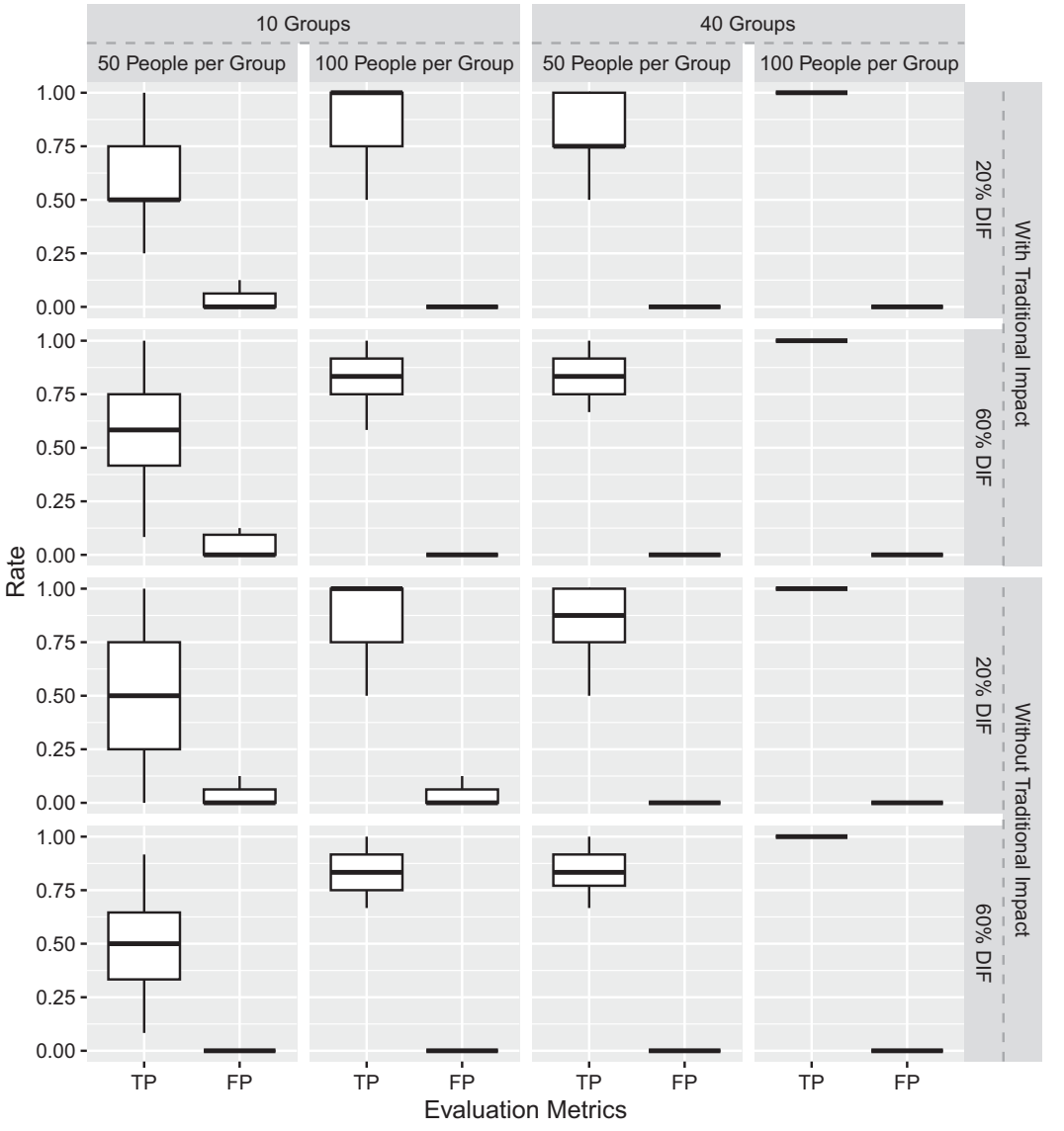


Figure 1. Simulation I results.

presence of traditional impact and a low proportion of DIF items. In general, this is consistent with prior studies, which found that it is more challenging to identify DIF effects on slope parameters than on intercepts (Bauer et al., 2020; Wang et al., 2023). Regarding the manipulated factors, while the number of groups and the sample size per group have consistent effects across Simulations I–III, the influence of DIF proportion and traditional impact becomes more pronounced in this study. Lastly, compared to Simulation I, DIF detection results for the intercept exhibit lower TP rates when $S = 10$. Additional guidance on the use of this method is provided in the Discussion section.

3.4. Simulation IV: UDIF and NUDIF detection with intersectional impact

The final simulation study evaluates the 2PL-RisM model, with the TP and FP rates summarized in Figure 4. In general, the method results in desirable TP and FP rates for detecting intersectional

https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press

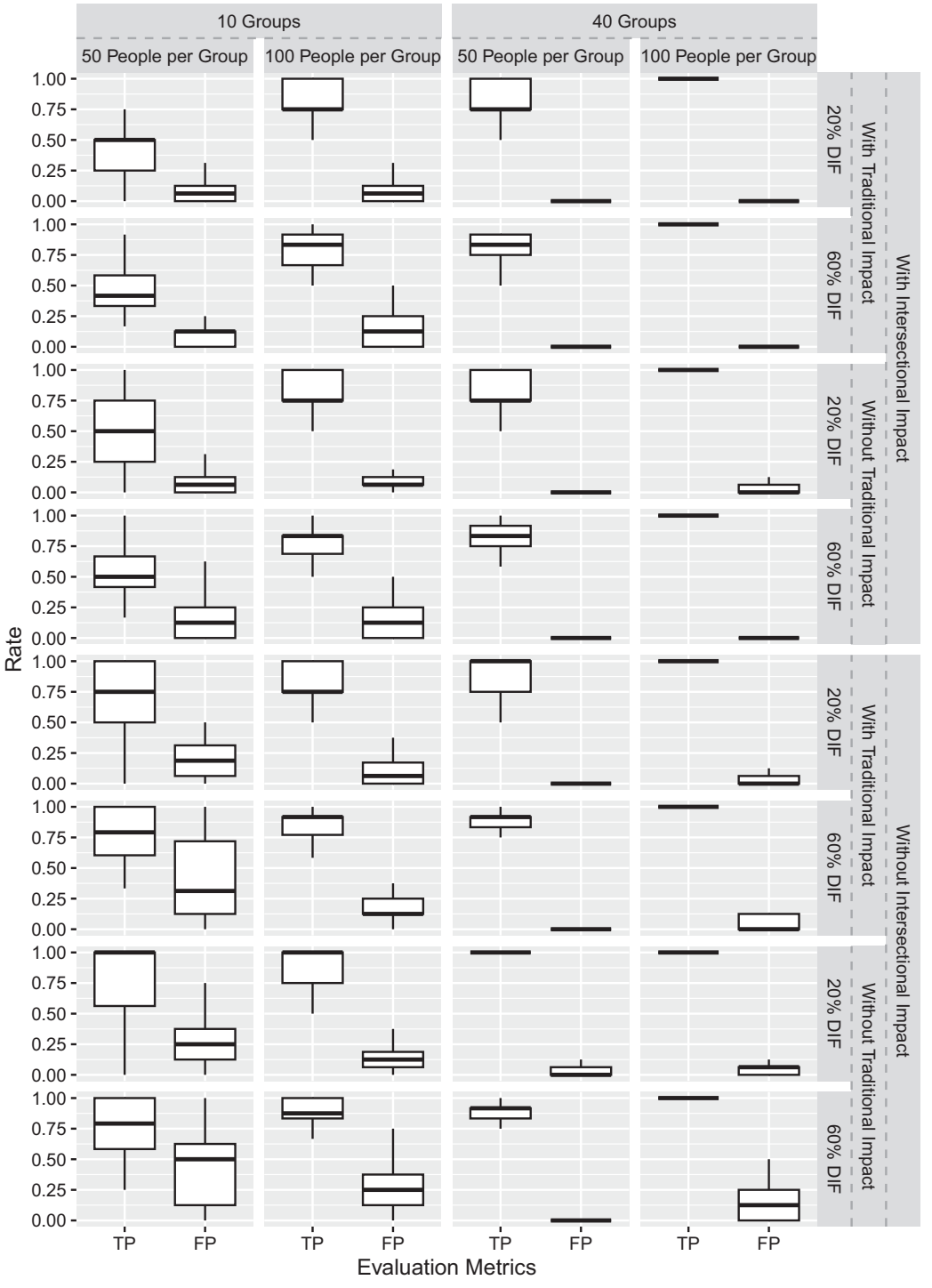
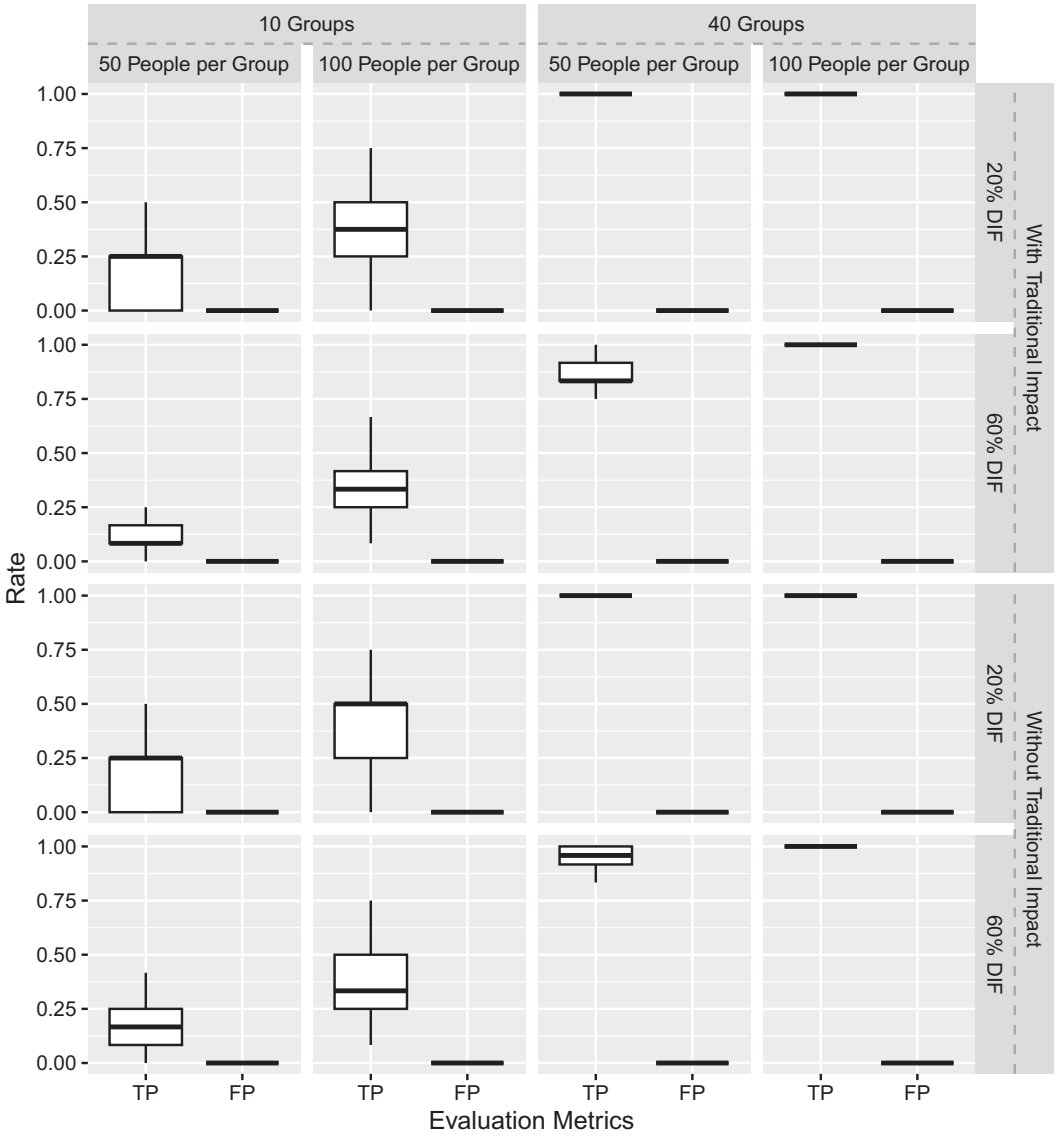


Figure 2. Simulation II results.

https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press



(a) *Intersectional NUDIF detection (on intercepts).*

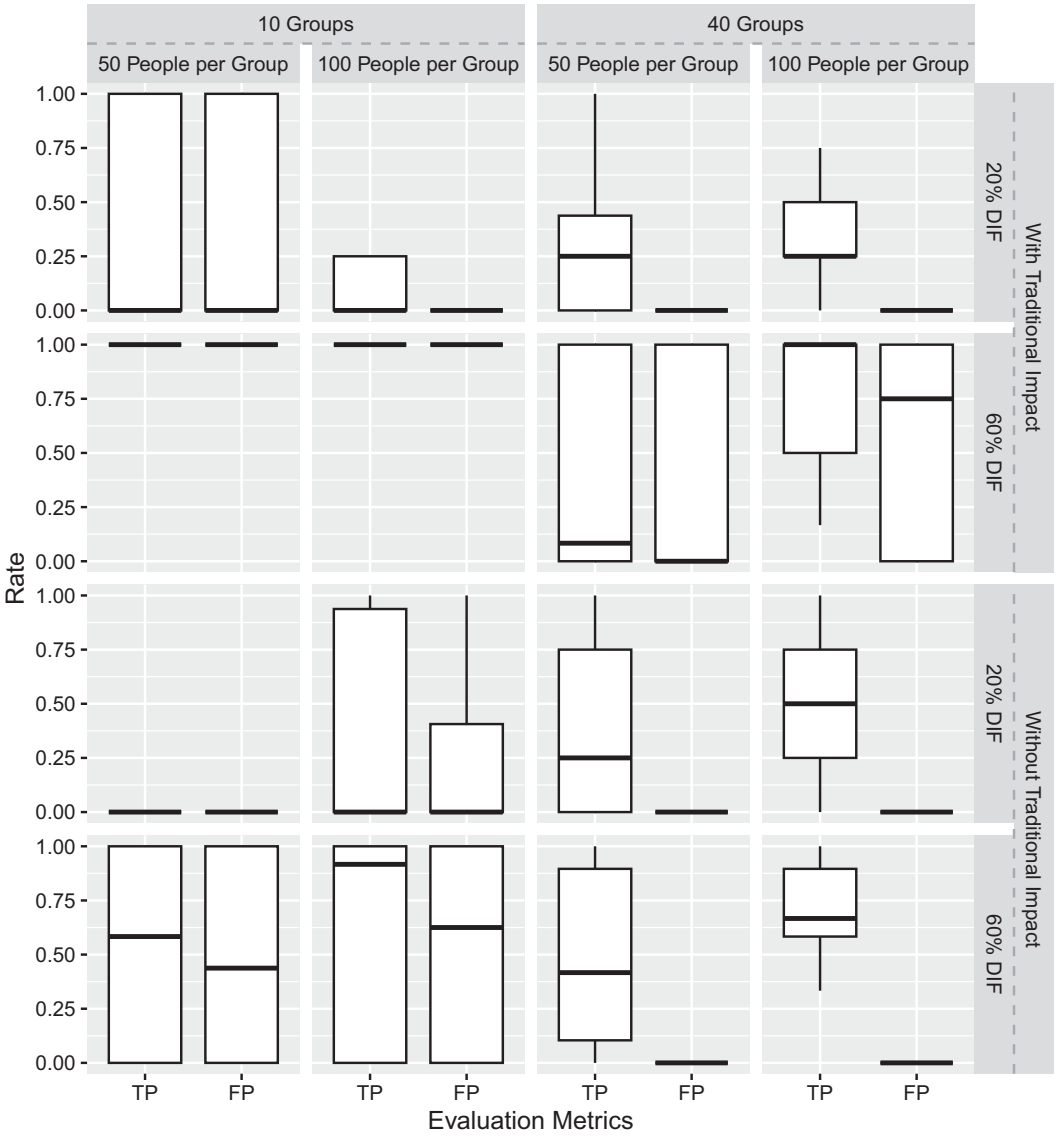
Figure 3. Simulation III results.

UDIF, but performs poorly in detecting intersectional NUDIF. In fact, the NUDIF detection results are generally unacceptable across nearly all conditions.

4. Empirical study

A real data set from the Programme for International Student Assessment (PISA) is used to demonstrate the performance of the four methods in this article. PISA is a well-known international large-scale assessment that tests the skills and knowledge of 15-year-old students in mathematics, reading, and science (OECD, 2019). In this study, we use a subset of the PISA 2018 science assessment, including

https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press



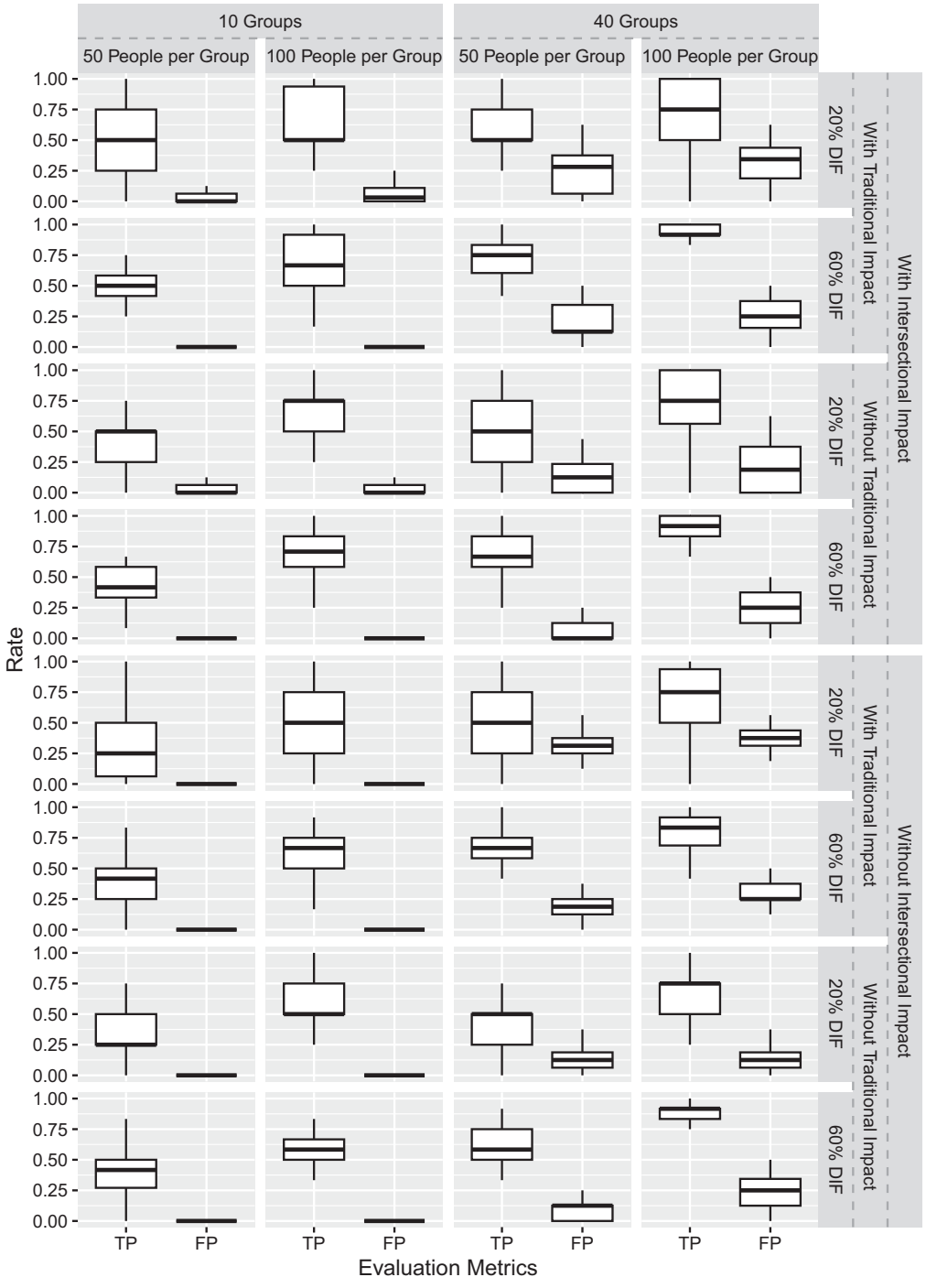
(b) *Intersectional NUDIF detection (on slopes).*

Figure 3. Continued.

dichotomous responses of 7,002 students on 19 items. Three demographic variables are considered: (1) country (eight countries in the subset), (2) sex (male or female), and (3) highest parental education (below or at least college level). These variables are chosen due to their frequent consideration in studies on educational equity. The full combination of these variables results in 32 intersectional groups, with their corresponding sample sizes summarized in Table 4.

Before discussing our empirical findings, we introduce a feasible way to tune the hyperparameter c in GIC (Lyu et al., 2025). Figure 5 illustrates the procedure, where c is plotted against J_{DIF}, the number of items exhibiting intersectional DIF. Similar to the scree plots in factor analysis, Figure 5 suggests that the models chosen by GIC with $c = 1.05$, which corresponds to the "elbow" of the plot. In practice, the choice of c can also depend on research goals. In certain high-stakes testing contexts, a higher FP rate

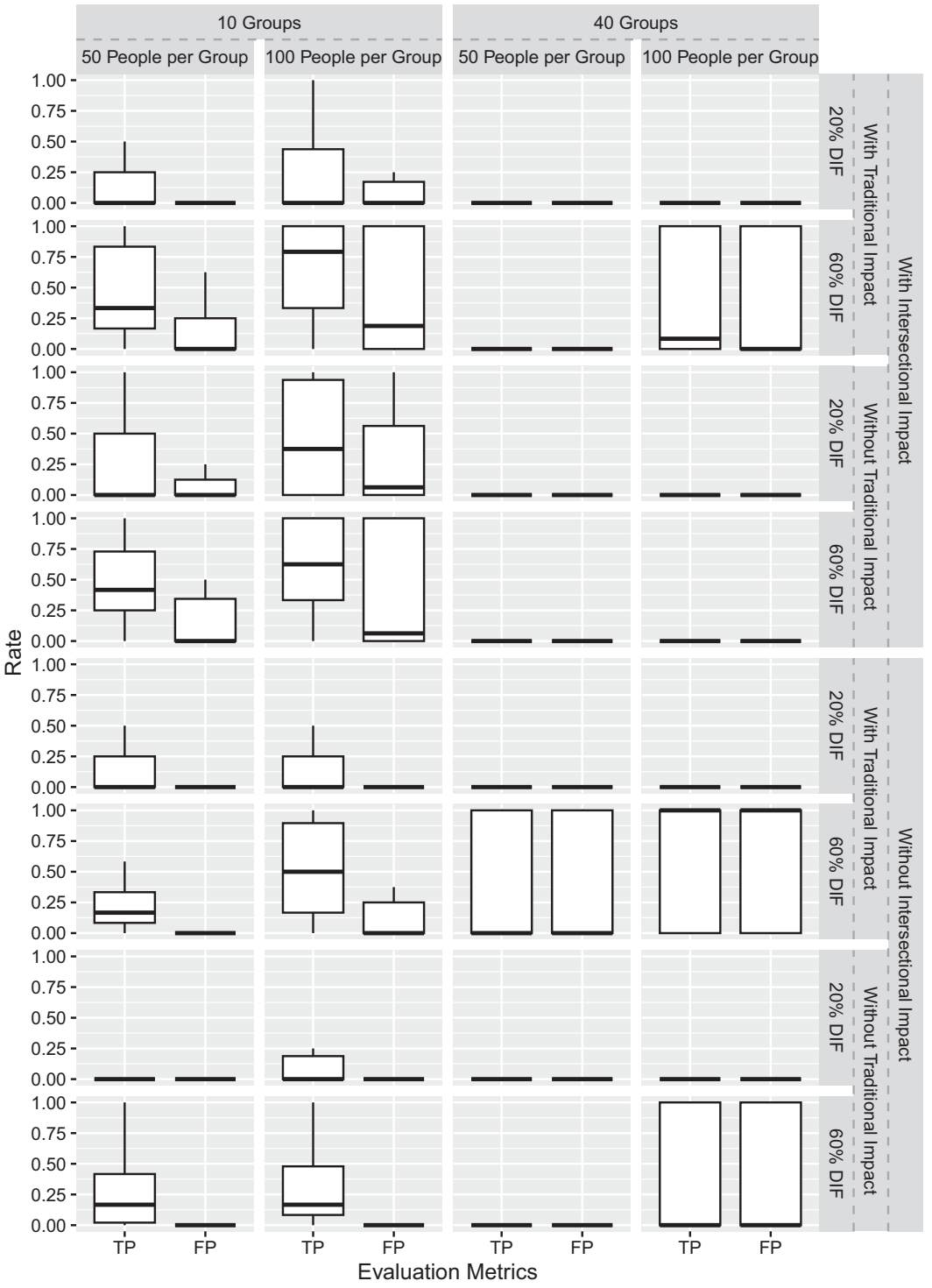
https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press



(a) *Intersectional NUDIF detection (on intercepts).*

Figure 4. Simulation IV results.

https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press



(b) *Intersectional NUDIF detection (on slopes).*

Figure 4. Continued.

https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press

Table 4. Sample size for each group in the empirical study

	Below college		At least college	
	Male	Female	Male	Female
ALB	175	120	93	94
ARE	300	233	951	927
AUS	258	282	536	522
AUT	110	133	184	189
BEL	115	98	255	293
BGR	71	62	100	111
BIH	68	68	53	86
BLR	38	35	205	237

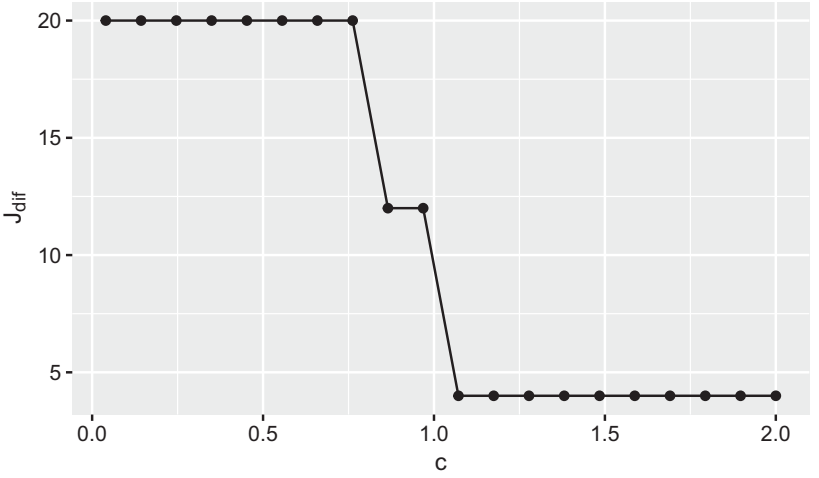


Figure 5. Relationship between the number of items exhibiting intersectional DIF and *c*.

Table 5. DIF detection results of the empirical study

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
2PL-Ri							*b*										*b*	*b*	*b*
2PL-RiM					*b*		*b*										*b*	*b*	*b*
2PL-Ris		*b*					*b*										*b*	*b*	*b*
2PL-RisM		*b*	*b*	*b*															

Note: *b* indicates UDIF, and NUDIF is not detected.

may be acceptable in order to achieve a high TP rate, as undetected DIF can lead to serious fairness concerns.

The empirical data set is analyzed using each of the four models, and the results are summarized in Table 5, where *a* and *b* refer to (intersectional) NUDIF and UDIF, respectively. Items 7, 17, 18, and 19 are flagged as UDIF items by most models, and no item is flagged as NUDIF.

Given the simulation results indicating unsatisfactory performance in detecting NUDIF, we focus our empirical analysis on the results from the 2PL-Ri and 2PL-RiM models. To validate the empirical

https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press

results, we estimate the RiM and Ri model using Markov chain Monte Carlo (MCMC) via the `rstan` package (Stan Development Team, 2024), allowing random item effects only for the consistently flagged items (i.e., items 7, 17, 18, and 19). For comparison, we also fit a constrained version of the model without random item effects. Model comparison based on the leave-one-out information criterion (LOOIC) reveals that the model with random item effects provides a significantly better fit, with a LOOIC difference of 107 and a standard error of 21.8. Furthermore, for comparison, a total score-based method is also employed to examine intersectional DIF (Belzak, 2023). Specifically, Belzak's method (2023) uses regularized logistic regression, with total score as the matching criterion and intersectionality modeled through interactions among demographic variables. This method is chosen due to its similarity to our proposed methods, as it accounts for both main and intersectional DIF effects with a primary focus on UDIF. However, this method has two limitations: (1) it does not automatically account for impact, since the total score is not directly regressed on demographic variables and (2) it may struggle with a large number of demographic variables, given that interactions are modeled using fixed effects. Despite these limitations, the method identifies UDIF in items 2, 5, 7, 14, 18, and 19, which largely aligns with the findings from our proposed methods.

5. Discussion

This study proposes a novel random effects IRT approach for detecting intersectional DIF and demonstrates the feasibility of applying a regularized GVEM algorithm in this context. By including both item-level and person-level random effects, the model accounts for intersectional DIF and impact effects arising from multiple demographic variables. Through the GVEM framework, all model parameters can be updated by closed-form solutions when detecting UDIF, resulting in a computationally efficient model estimation procedure. Simulation results show that the proposed methods can effectively detect UDIF. We have further extended the method to detect intersectional NUDIF, which is known to be more challenging. In this setting, all model parameters except the main and random effects on item discrimination (i.e., $\boldsymbol{\gamma}_j$ and $\bar{\sigma}^2_{a_j}$) have closed-form solutions (see the Appendix for details). The simulation results reveal that the number of groups has the most substantial impact on performance, followed by the sample size per group, the proportion of DIF items, and the presence or absence of impact. In terms of computational efficiency, the method performs well on standard hardware. On a laptop with an Intel i7-12700H CPU, the runtimes for a typical setting (i.e., 20 items, 20% DIF items, 40 groups, and 100 people per group) with a single regularization parameter range from 7.23 to 12.41 seconds, depending on the model used. These results underscore the scalability of the proposed approach for large-scale assessments.

In this study, intersectional DIF is modeled using random effects, and variation in group sizes may influence the methods' performance. Literature on multilevel modeling has shown that unequal cluster sizes can reduce both the power to detect true effects and the efficiency of estimating fixed and random components (Candel & Breukelen, 2009; Kush et al., 2022; Manatunga et al., 2001). Specifically, Candel & Breukelen (2009) found that the relative efficiency (RE) of the random intercept variance estimator can drop to between 84% and 95%, depending on the distribution and range of the cluster sizes. They also found that the loss in RE can be recovered by increasing the number of clusters, where the compensatory adjustment is given by $1/RE - 1$. For example, if the RE is 84%, then $1/0.84 - 1 \approx 0.19$, suggesting that an increase of 19% more clusters is needed to restore the original efficiency.

While we explore intersectional NUDIF detection alongside UDIF, the results for intersectional NUDIF detection are unsatisfactory, particularly when the model simultaneously accounts for intersectional impact. A supplementary simulation study demonstrates that even when response data are generated from the 2PL-Ris model, which includes intersectional NUDIF, the 2PL-Ri model still effectively identified items exhibiting intersectional UDIF. This suggests that, in practice, researchers should mainly rely on the detection results for intersectional UDIF when using the proposed methods. The results for intersectional NUDIF should be interpreted with caution and be used primarily in cases where intersectional impact is not included and when both the sample size and the number of groups

https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press

are sufficiently large. Despite these challenges, our framework provides a foundation for future advances in intersectional NUDIF detection. In a pilot study where only random effects, but not main effects, were considered, the methods demonstrate better results for NUDIF detection, suggesting potential for improvement. Future studies could explore the incorporation of regularization in both random effects and main effects to overcome estimation challenges. In addition, modeling slope parameters with lognormal distributions instead of truncated normal distributions may offer further improvement.

Another limitation of the simulation studies is that how demographic variables affect the variance of ability is not considered. Specifically, while mean latent traits are allowed to vary across groups, the within-group variance is assumed to be constant. Although this assumption aligns with most DIF research, several educational studies, such as Baye & Monseur (2016) and Gray et al. (2019), have discovered differences in latent trait variances among demographic groups. Future research may explore how demographic variables influence the variance of latent traits.

This study employs a variational approach to approximate the log marginal likelihood. Although parameter estimation may be biased due to the use of a mean-field Gaussian distribution family to approximate the posterior distribution of the latent variables, the approximation becomes increasingly accurate with larger sample sizes. Nevertheless, the estimation of latent variables, including group-specific item parameters with random effects and person abilities, may not be sufficiently accurate. To address this issue, we propose applying the standard MCMC procedure for their estimation, as described in the empirical study. Alternatively, future research could consider using best linear unbiased prediction (BLUP) as a complementary or alternative approach to MCMC.

Lastly, it is important to recognize that DIF arises within a complex social context. Each individual carries a unique set of experiences that shape their learning and life trajectories. However, when patterns of advantage or disadvantage emerge at the group level, they serve as a reminder that systemic discrimination continues to persist. Thus, while detecting DIF is a crucial first step in examining issues of fairness, it must be followed by deeper investigations into the underlying causes of structural inequality.

Data availability statement. The code that supports the findings of this study will be available on the project webpage (https://sites.uw.edu/pmetrics/projects/) shortly as we are still working on creating user-friendly R package and Shiny App. The real data was downloaded directly from the PISA website.

Author contributions. H.R. performed the formal analysis and initial draft writing; W.L. contributed to the derivation of the algorithms, C.W. contributed the original idea, method development, partial writing, reviewing and editing, and obtaining funding; G.X. contributed the original idea, method development, partial writing, reviewing and editing, and obtaining funding.

Funding statement. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D200015 and R305D240021 to University of Washington, National Science foundation, through grant EDU-CORE #2300382 to University of Washington, through grant SES-1846747, and SES-2150601 to University of Michigan, the research grant from Duolingo English Test, and the University of Washington BIRCH Center M-PARC Award. The opinions expressed are those of the authors and do not represent views of the Institute of Education Sciences or the U.S. Department of Education or National Science Foundation or Duolingo, Inc.

Competing interests. The authors declare no competing interests.

References

Adam, N. S., Twabi, H. S., & Manda, S. O. (2021). A simulation study for evaluating the performance of clustering measures in multilevel logistic regression. *BMC Medical Research Methodology*, *21*, 245. https://doi.org/10.1186/s12874-021-01417-4

Albano, T., French, B. F., & Vo, T. T. (2024). Traditional vs intersectional DIF analysis: Considerations and a comparison using state testing data. *Applied Measurement in Education*, *37*(1), 57–70. https://doi.org/10.1080/08957347.2024.2311935

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *The standards for educational and psychological testing*. The publisher is American Educational Research Association (AERA).

Baye, A., & Monseur, C. (2016). Gender differences in variability and extreme scores in an international context. *Large-Scale Assessments in Education*, *4*(1), 1. https://doi.org/10.1186/s40536-015-0015-x

Belzak, W. C. M. (2023). The multidimensionality of measurement bias in high-stakes testing: Using machine learning to evaluate complex sources of differential item functioning. *Educational Measurement: Issues and Practice*, *42*(1), 24–33. https://doi.org/10.1111/emip.12486

Belzak, W. C. M., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*, *25*(6), 673–690. https://doi.org/10.1037/met0000253

Bishop, C. M. (2006). *Pattern recognition and machine learning New York*. Springer.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773

Candel, M. J. J. M., & Breukelen, G. J. P. V. (2009). Varying cluster sizes in trials with clusters in one treatment arm: Sample size adjustments when testing treatment effects with linear mixed models. *Statistics in Medicine*, *28*(18), 2307–2324. https://doi.org/10.1002/sim.3620

Cho, A. E., Xiao, J., Wang, C., & Xu, G. (2024). Regularized variational estimation for exploratory item factor analysis. *Psychometrika*, *89*, 347–375. https://doi.org/10.1007/s11336-022-09874-6

Cho, A. E., Wang, C., Zhang, X., & Xu, G. (2021). Gaussian variational estimation for multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, *74*(S1), 52–85. https://doi.org/10.1111/bmsp.12219

Cole, E. R. (2009). Intersectionality and research in psychology. *American Psychologist*, *64*(3), 170–180. https://doi.org/10.1037/a0014564

Bauer, D. J., Belzak, W. C. M., & Cole, V. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning [PMID: 33132679]. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 43–55. https://doi.org/10.1080/10705511.2019.1642754

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*(4), 533–559. https://doi.org/10.1007/s11336-008-9092-x

Evans, C. R., Borrell, L. N., Bell, A., Holman, D., Subramanian, S., & Leckie, G. (2024). Clarifications on the intersectional MAIHDA approach: A conceptual guide and response to Wilkes and Karimi (2024). *Social Science and Medicine*, *350*, 116898. https://doi.org/10.1016/j.socscimed.2024.116898

Evans, C. R., Williams, D. R., Onnela, J.-P., & Subramanian, S. (2018). A multilevel approach to modeling health inequalities at the intersection of multiple social identities. *Social Science and Medicine*, *203*, 64–73. https://doi.org/10.1016/j.socscimed.2017.11.011

Fan, Y., & Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, *75*(3), 531–552. https://doi.org/10.1111/rssb.12001

Gray, H., Lyth, A., McKenna, C., Stothard, S., Tymms, P., & Copping, L. (2019). Sex differences in variability across nations in reading, mathematics and science: A meta-analytic extension of Baye and Monseur (2016). *Large-scale Assessments in Education*, *7*(1), 2. https://doi.org/10.1186/s40536-019-0070-9

Gu, Y., & Xu, G. (2019). Learning attribute patterns in high-dimensional structured latent attribute models. *Journal of Machine Learning Research*, *20*(115), 1–58. http://jmlr.org/papers/v20/19-197.html

Huang, Q., Bolt, D. M., & Lyu, W. (2024). Investigating item complexity as a source of cross-national DIF in TIMSS math and science. *Large-scale Assessments in Education*, *12*, 12. https://doi.org/10.1186/s40536-024-00200-3

Johfre, S. S., & Freese, J. (2021). Reconsidering the reference category. *Sociological Methodology*, *51*(2), 253–269. https://doi.org/10.1177/0081175020982632

Jong, M. G. D., Steenkamp, J.-B. E. M., & Fox, J.-P. (2007). Relaxing measurement invariance in crossnational consumer research using a hierarchical IRT model. *Journal of Consumer Research*, *34*(2), 260–278. https://doi.org/10.1086/518532

Keller, L., Lüdtke, O., Preckel, F., & Brunner, M. (2023). Educational inequalities at the intersection of multiple social categories: An introduction and systematic review of the multilevel analysis of individual heterogeneity and discriminatory accuracy (MAIHDA) approach. *Educational Psychology Review*, *35*, 31. https://doi.org/10.1007/s10648-023-09733-5

Kim, J., & Wilson, M. (2020). Polytomous item explanatory IRT models with random item effects: Concepts and an application. *Measurement*, *151*, 107062. https://doi.org/10.1016/j.measurement.2019.107062

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86. https://api.semanticscholar.org/CorpusID:120349231

Kush, J. M., Konold, T. R., & Bradshaw, C. P. (2022). Statistical power for randomized controlled trials with clusters of varying size. *The Journal of Experimental Education*, *90*(3), 673–692. https://doi.org/10.1080/00220973.2021.1873089

Lathrop, Q. N., & Cheng, Y. (2017). Item cloning variation and the impact on the parameters of response models. *Psychometrika*, *82*(1), 245–263. https://doi.org/10.1007/s11336-016-9513-1

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

Lyu, W., Wang, C., & Xu, G. (2025). Multi-group regularized Gaussian variational estimation: Fast detection of DIF. *Psychometrika*, *90*(1), 2–23. https://doi.org/10.1017/psy.2024.15

Ma, C., Ouyang, J., & Xu, G. (2023). Learning latent and hierarchical structures in cognitive diagnosis models. *Psychometrika*, *88*(1), 175–207. https://doi.org/10.1007/s11336-022-09867-5

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Manatunga, A. K., Hudgens, M. G., & Chen, S. (2001). Sample size estimation in cluster randomized studies with varying cluster size. *Biometrical Journal*, *43*(1), 75–86. https://doi.org/10.1002/1521-4036(200102)43:1<75::AID-BIMJ75>3.0.CO;2-N

Merlo, J. (2018). Multilevel analysis of individual heterogeneity and discriminatory accuracy (MAIHDA) within an intersectional framework. *Social Science and Medicine*, *203*, 74–80. https://doi.org/10.1016/j.socscimed.2017.12.026

Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, *7*, 34. https://doi.org/10.1186/1471-2288-7-34

Muthén, B., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociological Methods & Research*, *47*(4), 637–664. https://doi.org/10.1177/0049124117701488

Núñez, A.-M. (2014). Employing multilevel intersectionality in educational research: Latino identities, contexts, and college access. *Educational Researcher*, *43*(2), 85–92. https://doi.org/10.3102/0013189X14522320

OECD. (2019). *PISA 2018 results (volume I): What students know and can do*. OECD Publishing. https://doi.org/10.1787/5f07c754-en

Pastor, D. A. (2003). The use of multilevel item response theory modeling in applied research: An illustration. *Applied Measurement in Education*, *16*(3), 223–243. https://doi.org/10.1207/S15324818AME1603_4

Rijmen, F., & Jeon, M. (2013). Fitting an item response theory model with random item effects across groups by a variational approximation method. *Annals of Operations Research*, *206*, 647–662. https://doi.org/10.1007/s10479-012-1181-7

Russell, M., & Kaplan, L. (2021). An intersectional approach to differential item functioning: Reflecting configurations of inequality. *Practical Assessment Research & Evaluation*, *26*(21), 1–17.

Russell, M., Szendey, O., & Kaplan, L. (2021). An intersectional approach to DIF: Do initial findings hold across tests? *Educational Assessment*, *26*(4), 284–298. https://doi.org/10.1080/10627197.2021.1965473

Russell, M., Szendey, O., & Li, Z. (2022). An intersectional approach to DIF: Comparing outcomes across methods. *Educational Assessment*, *27*(2), 115–135. https://doi.org/10.1080/10627197.2022.2094757

Stan Development Team. (2024). RStan: The R interface to Stan [R package version 2.32.6]. https://mc-stan.org/

Sulis, I., & Toland, M. D. (2017). Introduction to multilevel item response theory analysis: Descriptive and explanatory models. *The Journal of Early Adolescence*, *37*(1), 85–128. https://doi.org/10.1177/0272431616642328

Swaminathan, H.,& Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370. http://www.jstor.org/stable/1434855

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, *80*(1), 21–43. https://doi.org/10.1007/s11336-013-9377-6

Wang, C. (2024). A diagnostic facet status model (DFSM) for extracting instructionally useful information from diagnostic assessment. *Psychometrika*, *89*(3), 747–773. https://doi.org/10.1007/s11336-024-09971-8

Wang, C., Zhu, R., & Xu, G. (2023). Using lasso and adaptive lasso to identify DIF in multidimensional 2PL models [PMID: 35086405]. *Multivariate Behavioral Research*, *58*(2), 387–407. https://doi.org/10.1080/00273171.2021.1985950

A. Appendix

The derivation of the 2PL-RisM model estimation procedure is shown below. The model is presented in Equations (1) and (2). Applying the local variational method in Equation (7), we obtain

$$
\begin{aligned}
&\int [\log p(\boldsymbol{Y},\boldsymbol{Z})]\, q(\boldsymbol{Z})\mathrm{d}\boldsymbol{Z} \\
=&\int [\log \mathbb{P}(\boldsymbol{Y}\mid\boldsymbol{Z}) + \log p(\boldsymbol{Z})]\, q(\boldsymbol{Z})\mathrm{d}\boldsymbol{Z} \\
\geq&\int \Bigg[\sum_{s=1}^{S}\sum_{i=1}^{N_s}\sum_{j=1}^{J}\Bigg\{ \log\frac{1}{1+\mathrm{e}^{-\xi_{ijs}}} + \left(y_{ijs}-\frac{1}{2}\right)(a_{js}\theta_{is}-b_{js}) - \frac{1}{2}\xi_{ijs} - \eta(\xi_{ijs})\left[(a_{js}\theta_{is}-b_{js})^2-\xi_{ijs}^2\right]\Bigg\} \\
&\quad -\frac{1}{2}\sum_{s=1}^{S}\Bigg\{ (N_s+2J+1)\log 2\pi + \left[\log\sigma_{\alpha_0}^2 + \frac{\alpha_{0s}^2}{\sigma_{\alpha_0}^2}\right] + \sum_{i=1}^{N_s}\left[\log\sigma_\theta^2 + \frac{(\theta_{is}-\alpha_{0s}-\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{X}_s)^2}{\sigma_\theta^2}\right] \\
&\qquad + \sum_{j=1}^{J}\left[\log\tilde{\sigma}_{a_j}^2 + \frac{(a_{js}-\tilde{\boldsymbol{\gamma}}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s)^2}{\tilde{\sigma}_{a_j}^2} + 2\log\Phi\left(\frac{\tilde{\boldsymbol{\gamma}}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s}{\tilde{\sigma}_{a_j}}\right) + \log\mathbb{1}\{a_{js}\geq 0\}\right] \\
&\qquad + \sum_{j=1}^{J}\left[\log\sigma_{b_j}^2 + \frac{(b_{js}-\boldsymbol{\beta}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s)^2}{\sigma_{b_j}^2}\right]\Bigg\}\Bigg] q(\boldsymbol{Z})\mathrm{d}\boldsymbol{Z} \\
\equiv&\int \mathcal{B}(\boldsymbol{Y},\boldsymbol{Z})q(\boldsymbol{Z})\mathrm{d}\boldsymbol{Z} \\
\equiv& Q(\boldsymbol{Y}).
\end{aligned}
\tag{A1}
$$

https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press

Then, the corresponding optimal variational distributions for the latent variables are

$$\begin{aligned} q_{\alpha_{0s}}(\alpha_{0s}) &\propto \exp \mathbb{E}_{\boldsymbol{Z}\setminus\{\alpha_{0s}\}}\left[\mathcal{B}(\boldsymbol{Y},\boldsymbol{Z})\right] \\ &\propto \exp \mathbb{E}_{\boldsymbol{Z}\setminus\{\alpha_{0s}\}}\left[-\frac{\alpha_{0s}^2}{2\sigma_{\alpha_0}^2}-\sum_{i=1}^{N_s}\frac{(\alpha_{0s}+\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{X}_s-\theta_{is})^2}{2\sigma_\theta^2}\right] \\ &\sim \mathcal{N}(\mu_{\alpha_{0s}},\sigma_{\alpha_{0s}}^2), \end{aligned}$$

where

$$\begin{aligned} \sigma_{\alpha_{0s}}^2 &= \frac{\sigma_{\alpha_0}^2\sigma_\theta^2}{N_s\sigma_{\alpha_0}^2+\sigma_\theta^2}, \\ \mu_{\alpha_{0s}} &= \frac{\sigma_{\alpha_0}^2}{N_s\sigma_{\alpha_0}^2+\sigma_\theta^2}\left(\sum_{i=1}^{N_s}\mu_{\theta_{is}}-N_s\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{X}_s\right); \end{aligned}$$

$$\begin{aligned} q_{\theta_{is}}(\theta_{is}) &\propto \exp \mathbb{E}_{\boldsymbol{Z}\setminus\{\theta_{is}\}}\left[\mathcal{B}(\boldsymbol{Y},\boldsymbol{Z})\right] \\ &\propto \exp \mathbb{E}_{\boldsymbol{Z}\setminus\{\theta_{is}\}}\left\{\sum_{j=1}^{J}\left[\left(y_{ijs}-\frac{1}{2}\right)a_{js}\theta_{is}-\eta(\xi_{ijs})(a_{js}\theta_{is}-b_{js})^2\right]\right. \\ &\qquad \left.-\frac{(\theta_{is}-\alpha_{0s}-\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{X}_s)^2}{2\sigma_\theta^2}\right\} \\ &\sim \mathcal{N}(\mu_{\theta_{is}},\sigma_{\theta_{is}}^2), \end{aligned}$$

where

$$\begin{aligned} \sigma_{\theta_{is}}^2 &= \frac{\sigma_\theta^2}{1+2\sigma_\theta^2\sum_{j=1}^{J}\eta(\xi_{ijs})(\mu_{a_{js}}^2+\sigma_{a_{js}}^2)}, \\ \mu_{\theta_{is}} &= \frac{\mu_{\alpha_{0s}}+\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{X}_s+\sigma_\theta^2\sum_{j=1}^{J}\mu_{a_{js}}\left[y_{ijs}-\frac{1}{2}+2\eta(\xi_{ijs})\mu_{b_{js}}\right]}{1+2\sigma_\theta^2\sum_{j=1}^{J}\eta(\xi_{ijs})(\mu_{a_{js}}^2+\sigma_{a_{js}}^2)}; \end{aligned}$$

$$\begin{aligned} q_{a_{js}}(a_{js}) &\propto \exp \mathbb{E}_{\boldsymbol{Z}\setminus\{a_{js}\}}\left[\mathcal{B}(\boldsymbol{Y},\boldsymbol{Z})\right] \\ &\propto \exp \mathbb{E}_{\boldsymbol{Z}\setminus\{a_{js}\}}\left\{\sum_{i=1}^{N_s}\left[\left(y_{ijs}-\frac{1}{2}\right)a_{js}\theta_{is}-\eta(\xi_{ijs})(a_{js}\theta_{is}-b_{js})^2\right]\right. \\ &\qquad \left.-\frac{(a_{js}-\tilde{\boldsymbol{\gamma}}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s)^2}{2\tilde{\sigma}_{a_j}^2}-\frac{\log \mathbb{1}\{a_{js}\geq 0\}}{2}\right\} \\ &\sim \mathcal{N}_+(\tilde{\mu}_{a_{js}},\tilde{\sigma}_{a_{js}}^2), \end{aligned}$$

where

$$\begin{aligned} \tilde{\sigma}_{a_{js}}^2 &= \frac{\tilde{\sigma}_{a_j}^2}{1+2\tilde{\sigma}_{a_j}^2\sum_{i=1}^{N_s}\eta(\xi_{ijs})(\mu_{\theta_{is}}^2+\sigma_{\theta_{is}}^2)}, \\ \tilde{\mu}_{a_{js}} &= \frac{\tilde{\boldsymbol{\gamma}}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s+\tilde{\sigma}_{a_j}^2\sum_{i=1}^{N_s}\left[y_{ijs}-\frac{1}{2}+2\eta(\xi_{ijs})\mu_{b_{js}}\right]\mu_{\theta_{is}}}{1+2\tilde{\sigma}_{a_j}^2\sum_{i=1}^{N_s}\eta(\xi_{ijs})(\mu_{\theta_{is}}^2+\sigma_{\theta_{is}}^2)}, \\ \sigma_{a_{js}}^2 &= \tilde{\sigma}_{a_{js}}^2\left\{1-\frac{\tilde{\mu}_{a_{js}}}{\sqrt{2\pi}\tilde{\sigma}_{a_{js}}\Phi\left(\frac{\tilde{\mu}_{a_{js}}}{\tilde{\sigma}_{a_{js}}}\right)}\exp\left(-\frac{\tilde{\mu}_{a_{js}}^2}{2\tilde{\sigma}_{a_{js}}^2}\right)-\frac{1}{2\pi\left[\Phi\left(\frac{\tilde{\mu}_{a_{js}}}{\tilde{\sigma}_{a_{js}}}\right)\right]^2}\exp\left(-\frac{\tilde{\mu}_{a_{js}}^2}{\tilde{\sigma}_{a_{js}}^2}\right)\right\}, \\ \mu_{a_{js}} &= \tilde{\mu}_{a_{js}}+\frac{\tilde{\sigma}_{a_{js}}}{\sqrt{2\pi}\Phi\left(\frac{\tilde{\mu}_{a_{js}}}{\tilde{\sigma}_{a_{js}}}\right)}\exp\left(-\frac{\tilde{\mu}_{a_{js}}^2}{2\tilde{\sigma}_{a_{js}}^2}\right); \end{aligned}$$

$$\begin{aligned} q_{b_{js}}(b_{js}) &\propto \exp \mathbb{E}_{\boldsymbol{Z}\setminus\{b_{js}\}}\left[\mathcal{B}(\boldsymbol{Y},\boldsymbol{Z})\right] \\ &\propto \exp \mathbb{E}_{\boldsymbol{Z}\setminus\{b_{js}\}}\left\{-\sum_{i=1}^{N_s}\left[\left(y_{ijs}-\frac{1}{2}\right)b_{js}+\eta(\xi_{ijs})(b_{js}-a_{js}\theta_{is})^2\right]-\frac{(b_{js}-\boldsymbol{\beta}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s)^2}{2\sigma_{b_j}^2}\right\} \\ &\sim \mathcal{N}(\mu_{b_{js}},\sigma_{b_{js}}^2), \end{aligned}$$

https://doi.org/10.1017/psy.2025.10046 Published online by Cambridge University Press

where

$$\sigma^2_{b_{js}} = \frac{\sigma^2_{b_j}}{1+2\sigma^2_{b_j}\sum_{i=1}^{N_s}\eta(\xi_{ijs})},$$
$$\mu_{b_{js}} = \frac{\boldsymbol{\beta}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s - \sigma^2_{b_j}\sum_{i=1}^{N_s}\left[y_{ijs}-\frac{1}{2}-2\eta(\xi_{ijs})\mu_{a_{js}}\mu_{\theta_{is}}\right]}{1+2\sigma^2_{b_j}\sum_{i=1}^{N_s}\eta(\xi_{ijs})}.$$

Given the optimal variational distributions derived above and the mean-field variational assumption, the expectation (i.e., the integral) in Equation (A1) can be computed as

$$\begin{aligned}
Q(\boldsymbol{Y}) = &\sum_{s=1}^{S}\sum_{i=1}^{N_s}\sum_{j=1}^{J}\left\{\log\frac{1}{1+\mathrm{e}^{-\xi_{ijs}}} + \left(y_{ijs}-\frac{1}{2}\right)\left(\mu_{a_{js}}\mu_{\theta_{is}}-\mu_{b_{js}}\right) - \frac{1}{2}\xi_{ijs}\right.\\
&\left. - \eta(\xi_{ijs})\left[\left(\mu^2_{a_{js}}+\sigma^2_{a_{js}}\right)\left(\mu^2_{\theta_{is}}+\sigma^2_{\theta_{is}}\right) - 2\mu_{a_{js}}\mu_{\theta_{is}}\mu_{b_{js}} + \mu^2_{b_{js}}+\sigma^2_{b_{js}} - \xi^2_{ijs}\right]\right\}\\
&- \frac{1}{2}\sum_{s=1}^{S}\left\{\left(N_s+2J+1\right)\log 2\pi + \left[\log\sigma^2_{\alpha_0} + \frac{\mu^2_{\alpha_{0s}}+\sigma^2_{\alpha_{0s}}}{\sigma^2_{\alpha_0}}\right]\right.\\
&\quad + \sum_{i=1}^{N_s}\left[\log\sigma^2_{\theta} + \frac{\sigma^2_{\theta_{is}}+\sigma^2_{\alpha_{0s}}+\left(\mu_{\theta_{is}}-\mu_{\alpha_{0s}}-\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{X}_s\right)^2}{\sigma^2_{\theta}}\right]\\
&\quad + \sum_{j=1}^{J}\left[\log\tilde{\sigma}^2_{a_j} + \frac{\sigma^2_{a_{js}}+\left(\mu_{a_{js}}-\tilde{\boldsymbol{\gamma}}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s\right)^2}{\tilde{\sigma}^2_{a_j}} + 2\log\Phi\left(\frac{\tilde{\boldsymbol{\gamma}}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s}{\tilde{\sigma}_{a_j}}\right)\right]\\
&\quad \left. + \sum_{j=1}^{J}\left[\log\sigma^2_{b_j} + \frac{\sigma^2_{b_{js}}+\left(\mu_{b_{js}}-\boldsymbol{\beta}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s\right)^2}{\sigma^2_{b_j}}\right]\right\}.
\end{aligned}$$

The objective function with log penalty is

$$Q'(\boldsymbol{Y}) = Q(\boldsymbol{Y}) - \lambda\sum_{j=1}^{J}\left(\log\tilde{\sigma}^2_{a_j} + \log\sigma^2_{b_j}\right).$$

By setting the derivative of the objective function with respect to each model parameter to be zero, we get the following parameters update rules:

$$\xi^2_{ijs} = \left(\mu^2_{a_{js}}+\sigma^2_{a_{js}}\right)\left(\mu^2_{\theta_{is}}+\sigma^2_{\theta_{is}}\right) - 2\mu_{a_{js}}\mu_{\theta_{is}}\mu_{b_{js}} + \mu^2_{b_{js}}+\sigma^2_{b_{js}},$$
$$\sigma^2_{\alpha_0} = \frac{1}{S}\sum_{s=1}^{S}\left(\sigma^2_{\alpha_{0s}}+\mu^2_{\alpha_{0s}}\right),$$
$$\sigma^2_{\theta} = \frac{1}{\sum_{s=1}^{S}N_s}\sum_{s=1}^{S}\left\{N_s\left[\sigma^2_{\alpha_{0s}} + \left(\mu_{\alpha_{0s}}+\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{X}_s\right)^2\right] + \sum_{i=1}^{N_s}\left[\sigma^2_{\theta_{is}} + \mu^2_{\theta_{is}} - 2\mu_{\theta_{is}}\left(\mu_{\alpha_{0s}}+\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{X}_s\right)\right]\right\},$$
$$\boldsymbol{\alpha}_1 = \left(\sum_{s=1}^{S}\frac{N_s}{\sigma^2_{\theta}}\boldsymbol{X}_s\boldsymbol{X}_s^{\mathrm{T}}\right)^{-1}\left(\sum_{s=1}^{S}\frac{\sum_{i=1}^{N_s}\mu_{\theta_{is}} - N_s\mu_{\alpha_{0s}}}{\sigma^2_{\theta}}\boldsymbol{X}_s\right),$$
$$\left(\tilde{\boldsymbol{\gamma}}_j, \tilde{\sigma}^2_{a_j}\right) = \underset{\left(\tilde{\boldsymbol{\gamma}}_j, \tilde{\sigma}^2_{a_j}\right)}{\operatorname{argmin}}\sum_{s=1}^{S}\left\{\log\tilde{\sigma}^2_{a_j} + \frac{\sigma^2_{a_{js}}+\left(\mu_{a_{js}}-\tilde{\boldsymbol{\gamma}}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s\right)^2}{\tilde{\sigma}^2_{a_j}} + 2\log\Phi\left(\frac{\tilde{\boldsymbol{\gamma}}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s}{\tilde{\sigma}_{a_j}}\right)\right\} + 2\lambda\log\tilde{\sigma}^2_{a_j},$$
$$\boldsymbol{\beta}_j = \left(\sum_{s=1}^{S}\tilde{\boldsymbol{X}}_s\tilde{\boldsymbol{X}}_s^{\mathrm{T}}\right)^{-1}\left(\sum_{s=1}^{S}\mu_{b_{js}}\tilde{\boldsymbol{X}}_s\right),$$
$$\sigma^2_{b_j} = \frac{1}{S+2\lambda}\sum_{s=1}^{S}\left[\sigma^2_{b_{js}} + \left(\mu_{b_{js}}-\boldsymbol{\beta}_j^{\mathrm{T}}\tilde{\boldsymbol{X}}_s\right)^2\right].$$

Cite this article: Ren, H., Lyu, W., Wang, C. and Xu, G. (2025). A Novel Method for Detecting Intersectional DIF: Multilevel Random Item Effects Model with Regularized Gaussian Variational Estimation. *Psychometrika*, 1–25. https://doi.org/10.1017/psy.2025.10046