This is a "preproof" accepted article for *Psychometrika*.
This version may be subject to change during the production process.
DOI: 10.1017/psy.2025.10072

REDUCING DIFFERENTIAL ITEM FUNCTIONING VIA PROCESS DATA

Ling Chen*, Susu Zhang†, Jingchen Liu*

*department of statistics, columbia university & †department of psychology, university of illinois urbana-champaign

December 1, 2025

Correspondence should be sent to Jingchen Liu,

E-Mail: jcliu@stat.columbia.edu
Phone:
Fax:
Website:

REDUCING DIFFERENTIAL ITEM FUNCTIONING VIA PROCESS DATA

Abstract

Test fairness is a major concern in psychometric and educational research. A typical approach for ensuring test fairness is through differential item functioning (DIF) analysis. DIF arises when a test item functions differently across subgroups that are typically defined by the respondents' demographic characteristics. Most of the existing research focuses on the statistical detection of DIF, yet less attention has been given to reducing or eliminating DIF. Simultaneously, the use of computer-based assessments has become increasingly popular. The data obtained from respondents interacting with an item are recorded in computer log files and are referred to as process data. Process data provide valuable insights into respondents' problem-solving strategies and progress, offering new opportunities for DIF analysis. In this paper, we propose a novel method within the framework of generalized linear models (GLMs) that leverages process data to reduce and understand DIF. Specifically, we construct a nuisance trait surrogate with the features extracted from process data. With the constructed nuisance trait, we introduce a new scoring rule that incorporates respondents' behaviors captured through process data on top of the target latent trait. We demonstrate the efficiency of our approach through extensive simulation experiments and an application to thirteen Problem Solving in Technology-Rich Environments (PSTRE) items from the 2012 Programme for the International Assessment of Adult Competencies (PIAAC) assessment.

Key words: differential item functioning, process data, item response theory, scoring rule

1. Introduction

Ensuring test fairness in educational and psychometric assessments has been a major concern for researchers. A typical approach to ensuring test fairness is through differential item functioning (DIF, Holland and Wainer, 2012) analysis. DIF occurs when an item's response function depends on not only the target latent trait to be measured by the item, but also the respondents' group memberships that are often linked to their demographic characteristics. When DIF is present, the measurement properties of the item differ systematically across groups, leading to measurement bias (Millsap, 2012).

The research on DIF predominantly focuses on the statistical detection of its existence. DIF detection with manifest groups is typically categorized as non-parametric (Holland and Thayer, 1986; Dorans and Kulick, 1986; Dorans and Holland, 1992; Zwick et al., 2000; Woods et al., 2013; Cao et al., 2017) and parametric (Lord, 1977, 1980; Rudner et al., 1980; Raju, 1988; Swaminathan and Rogers, 1990; Thissen et al., 2013). When the comparison groups are unavailable, latent DIF analysis compares how the item functions different across latent groups (Cho and Cohen, 2010; De Boeck et al., 2011; Cho et al., 2016b). More recently, DIF detection methods that do not require predefined group memberships or anchor items have been proposed (Chen et al., 2023; Wallin et al., 2024; Halpin, 2024; Ouyang et al., 2025). While research has explored methods on handling items with DIF (Cho et al., 2016a; Liu and Jane Rogers, 2022), items identified with significant DIF are often removed during item calibration, leading to wasted resources and efforts in their development and administration. Consequently, there has been growing interest in reducing or eliminating DIF, as well as understanding the underlying reasons for why DIF occurs (Ackerman and Ma, 2024).

One way to attribute the cause of DIF is multidimensionality, where DIF arises due to the presence of secondary dimensions in the latent space (Kok, 1988; Ackerman, 1992; Shealy and Stout, 1993). Ideally, differences in the response probabilities solely reflect variations in the latent ability that the item is designed to assess, which is the primary dimension. Secondary latent traits with heterogeneous distributions across subpopulations may also contribute to these differences. These secondary dimensions are called auxiliary if they are intentionally measured by

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

the item, or nuisance otherwise (Roussos and Stout, 1996). In our examples, computer proficiency is one of such nuisance traits correlated with problem solving strategies, response time, and the final responses. As we will show in the PIAAC data analysis, age is one of the background variables having substantial DIF for multiple items, part of which is due to differences in computer proficiency among age groups.

The multidimensional IRT (MIRT) model has been used to analyze DIF, where both the target trait and the nuisance trait are used to model item response probabilities (Ackerman, 1992; Shan and Xu, 2024; Wang et al., 2023a; Ackerman and Ma, 2024). Latent DIF models have also been used to investigate the secondary dimensions, using mixture models to identify latent groups as the secondary dimension and associating it with examinees' demographic characteristics (Cohen and Bolt, 2005; De Boeck et al., 2011). The multiple-indicator multiple-cause (MIMIC) model provides another approach from the dimensionality perspective, although only the primary dimension is used (De Boeck et al., 2011). In a mediated MIMIC model proposed in Cheng et al. (2016), a secondary dimension construct (the scale of self-confidence) is used as a potential mediator. Despite these advances, studies that rely only on response data face challenges, particularly when prior knowledge of nuisance traits is limited. Thus, it is of interest to identify the secondary dimensions from data sources beyond the response outcome data.

One data source that presents new opportunities for identifying secondary latent dimensions in DIF analysis is process data (He et al., 2021; Wang et al., 2023b; Li et al., 2024). Process data capture the problem-solving processes as respondents interact with computer-based test items. The respondents' actions are logged as time-stamped action sequences in computer log files, making process data a detailed record of respondents' behaviors. Three prominent examples of process data are from the Programme for International Student Assessment (PISA; e.g., OECD, 2012c), the Programme for the International Assessment of Adult Competencies (PIAAC; e.g., OECD, 2012a), and the National Assessment of Educational Progress (NAEP; e.g., Bergner and von Davier, 2019). These assessments not only measure skills traditionally tested with paper-and-pencil methods but also evaluate more complex abilities such as problem-solving in technology-rich environments. Compared to traditional outcome data that are typically dichotomous (correct/incorrect) or polytomous (partial credit), process data provides more

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

comprehensive information about the respondents' behaviors towards completing tasks. Process data have been proven useful for accurate assessment (Zhang et al., 2023), process-incorporated measurement models (Chen, 2020; Xiao et al., 2021; Liang et al., 2023; Xiao and Liu, 2024; Tang, 2024), strategy and behavioral pattern analysis (Gao et al., 2022; He et al., 2021; Ulitzsch et al., 2022; He et al., 2023), etc. Process data is potentially a good data source to identify nuisance traits. For example, engagement is an acknowledged nuisance trait (Wise and Kong, 2005; Wise and DeMars, 2006), and is usually measured by total response time or total number of actions (Sahin and Colvin, 2020). The first principal component of features extracted from the 2012 PIAAC process data is highly correlated with engagement (Tang et al., 2020a). If we are able to identify the secondary dimensions that lead to DIF, we could reduce or potentially remove DIF by appropriately incorporating such a dimension in the scoring rule.

In this paper, we propose a novel method for reducing DIF that leverages process data and introduce the corresponding scoring rule that only depends on the respondents' behaviors. We attribute DIF to multidimensionality and discuss the method within the framework of generalized linear models (GLM). We assume there is a reasonably accurate estimate of the target latent trait (primary dimension). The nuisance trait is partially predictable by the process data. The key innovation of our approach lies in constructing a surrogate for the nuisance trait using features extracted from process data. This surrogate is formulated as a linear combination of the process data features, with the weights determined by minimizing the maximum likelihood difference between models with and without the grouping variable. The motivation is to construct a measurement model with the target trait and nuisance trait surrogate, while minimizing the impact of the grouping variable on the measurement model. We show that in the simple case of linear model (classical test theory), the proposed optimization problem has a closed-form solution. For generalizes linear models, the optimization can be solved by well established numerical methods. We propose a new scoring rule that incorporates both the target latent trait and nuisance trait surrogate that reduces DIF. We stress that the scoring rule is purely based on respondents' responses. We further propose an iterative method to reduce the dependence on the initial target trait estimate. The effectiveness of this method is demonstrated through simulation studies and a case study.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

The rest of the paper is structured as follows. Section 2 outlines the methodology of the proposed approach, while Section 3 presents the results of the simulation experiments. In Section 4, we demonstrate a case study using the PIAAC 2012 dataset. Finally, we provide a discussion in Section 5.

2. Method

Consider N independent respondents and their process and outcome responses to one item of interest. Let $Y_i \in \{0, \ldots, C-1\}$ represent the response of the i-th respondent, where C is the number of possible responses. For example, when $C = 2$, $Y_i = 1$ indicates correct response and $Y_i = 0$ indicates otherwise. We use $\mathbf{Y} = (Y_1, \ldots, Y_N)^\top \in \mathbb{R}^N$ to denote the vector of responses from all N respondents. We introduce one grouping variable $Z_i \in \{0, 1\}$, where $Z_i = 0$ represents the reference group and $Z_i = 1$ represents the focal group. The item is assumed to measure a unidimensional latent trait, denoted by θ_i for respondent $i \in [N]$. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)^\top$ be the collection of latent traits of all respondents. Additionally, we extract process features $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\top \in \mathbb{R}^{N \times K}$ from the action sequences of the process data with the multidimensional scaling procedure proposed in Tang et al. (2020a), where $\mathbf{x}_i \in \mathbb{R}^K$. Since process features capture most of the useful information of process data when the feature dimension is sufficiently large (Tang et al., 2020a,b), we will use process features as a proxies for the original process data in this work.

2.1. Differential Item Functioning

We adopt a multidimensionality-based DIF framework. In addition to the target trait dimension, θ_i, a nuisance trait dimension, η_i, also influences the probability of the response. Denote $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_N)^\top$ as the vector encompassing the nuisance traits for all respondents. DIF occurs when the item response probability depends on the nuisance trait, and there is distributional difference of the nuisance trait among the two groups. Conditional on θ_i, η_i and Z_i, it is assumed that Y_i are independently distributed. When the distribution of η_i conditional on θ_i

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

differs across the two subgroups, we have

$$p(Y_i = y|\theta_i, Z_i = 0) = \int p(Y_i = y|\theta_i, \eta_i) \cdot p(\eta_i|\theta_i, Z_i = 0) d\eta_i$$
$$\neq \int p(Y_i = y|\theta_i, \eta_i) \cdot p(\eta_i|\theta_i, Z_i = 1) d\eta_i = p(Y_i = y|\theta_i, Z_i = 1).$$

Therefore, under this multidimensional framework, a uni-dimensional measurement model leads to DIF.

Specifically, we consider a GLM with the following conditional distribution:

$$Y_i \sim p(y|\mu_i), \quad \text{where} \;\; g(\mu_i) = d + a_0\theta_i + a_1\eta_i + \lambda Z_i. \tag{1}$$

Here $g(\cdot)$ is the link function with respect to the response mean $\mu_i = \mathbb{E}[Y_i|\theta_i, \eta_i, Z_i]$, and $d, \boldsymbol{a} := (a_0, a_1), \lambda$ are unknown coefficients. When $\lambda = 0$, the distributional difference of $\eta_i|\theta_i$ among the two subgroups is the only DIF source in a uni-dimensional measurement model. The model specified by (1) is quite general, as we allow $g(\cdot)$ to take a general form for a wide range of response types such as binary, polytomous, and continuous responses. For binary responses, the logistic regression model with $g(\mu_i) = \ln(\mu_i/(1-\mu_i))$ is referred to as the multidimensional two-parameter logistic (M2PL) IRT model. In addition, $g(\mu_i) = \Phi^{-1}(\mu_i)$ corresponds to the probit regression model with $\Phi(\cdot)$ being the cumulative distribution function of the standard normal distribution. When $g(\mu_i) = \mu_i$, the model becomes a linear regression model for continuous responses.

2.2. DIF Reduction

One of the primary challenges in applying the multidimensionality-based DIF analysis is that the nuisance trait, η_i, is unobserved and is generally difficult to be directly measured. We propose to construct a surrogate for the unobserved nuisance trait and incorporate it in the item response function to correct DIF. Specifically, we aim to build a surrogate, $\widehat{\eta}_i$, such that DIF is only attributed to the distributional differences in $\widehat{\eta}_i|\theta_i$, and the conditional probabilities of item responses become approximately equal across different subgroups. Formally, we achieve the following:

$$p(Y_i = y|\theta_i, \widehat{\eta}_i, Z_i = 0) \approx p(Y_i = y|\theta_i, \widehat{\eta}_i, Z_i = 1). \tag{2}$$

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

We propose using process data to construct a nuisance trait surrogate, driven by two key motivations. Firstly, process data capture the entire sequence of actions taken by each respondent as they interact with and solve an item, providing a rich source of information on various nuisance traits. Second, process data typically predict the final response with perfect accuracy. By analyzing the respondent's full sequence of actions, we can infer whether they answered the item correctly or their partial scores. In theory, adding all available process features in the model eliminates DIF. Yet this approach is not ideal as the resulting measurement model solely depends on the process features and provides little information of the target trait.

The core of our proposed method is to identify an optimal linear combination of process features as a surrogate for the nuisance trait η_i. Specifically, we consider

$$\eta_i = \boldsymbol{\omega}^\top \mathbf{x}_i,$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)^\top \in \mathbb{R}^K$ is the weight vector and we assume $\|\boldsymbol{\omega}\| = 1$ for model identifiability. Our objective in identifying $\boldsymbol{\omega}$ is to minimize a quantity equivalent to the likelihood ratio test statistic,

$$L(\boldsymbol{\omega}) := \max_{d,\boldsymbol{a},\lambda} l(d, \boldsymbol{a}, \lambda) - \max_{d,\boldsymbol{a},\lambda=0} l(d, \boldsymbol{a}, \lambda), \tag{3}$$

where $l(\cdot)$ is the log-likelihood function,

$$l(d, \boldsymbol{a}, \lambda) = \sum_{i=1}^{N} \log p(Y_i | \theta_i, \eta_i, Z_i). \tag{4}$$

Function (3) quantifies how much model fit is increased after adding the grouping variable into the model, thus can be viewed as a quantification of the DIF effect. When there is no DIF exhibited, adding the grouping variable into the model would barely increase the likelihood, and we would expect $L(\boldsymbol{\omega})$ to be close to 0. This objective function enables us to optimize $\boldsymbol{\omega}$ by comparing models with and without the grouping variable, ultimately reducing or removing DIF. Therefore, we propose the estimation of $\boldsymbol{\omega}$ to be the minimizer of the objective function

$$\widehat{\boldsymbol{\omega}} = \arg \min_{\|\boldsymbol{\omega}\|=1} L(\boldsymbol{\omega}). \tag{5}$$

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

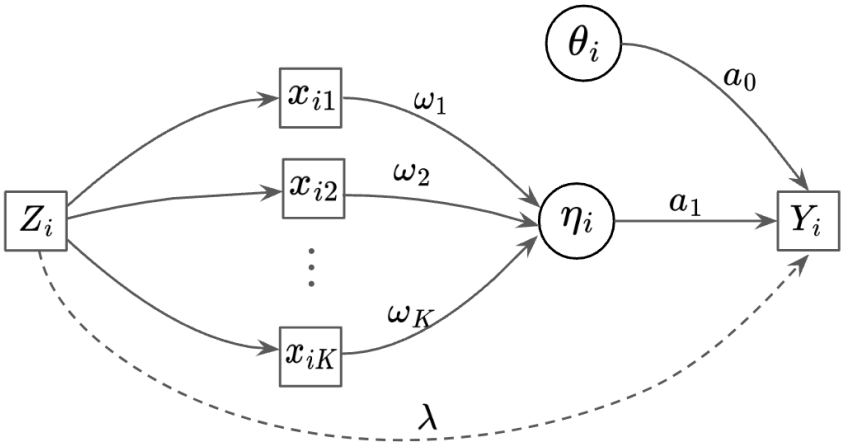

FIGURE 1.
Measurement model without the intercept.

and the nuisance trait surrogate is $\widehat{\eta}_i = \widehat{\boldsymbol{\omega}}^\top \mathbf{x}_i$. We will show that the solution of (5) has a closed-form solution in the linear regression model with one grouping variable; see Section 2.3. In other cases, we can optimize the objective function numerically.

Figure 1 displays the updated measurement model incorporating both the nuisance trait and target trait. With the nuisance trait surrogate, we update the initial estimate of the target trait with the two-dimensional measurement model. Consider the case with J items, among which items $j \in \mathcal{B} \subset [J]$ exhibit DIF. Suppose the nuisance trait surrogates $\widehat{\boldsymbol{\eta}}_j = (\widehat{\eta}_{ij}) \in \mathbb{R}^N, j \in \mathcal{B}$ have been estimated for the DIF items, then we obtain the MLE of the item parameters $\widehat{d}_j, \widehat{a}_{0j}, \widehat{a}_{1j}$ with λ_j fixed as 0 in (1) for $j \in \mathcal{B}$, and $\widehat{d}_j, \widehat{a}_{0j}$ with a_{1j}, λ_j fixed as 0 for $j \in [J]\backslash\mathcal{B}$. The target trait estimate is then updated by the maximum likelihood estimate (MLE)

$$\widehat{\theta}_i = \underset{\theta}{\operatorname{argmax}} \sum_{j=1}^{J} \log p_{ij}(\theta), \tag{6}$$

where

$$p_{ij}(\theta) = p(Y_{ij}|\theta, \widehat{\eta}_{ij}, \widehat{d}_j, \widehat{a}_{0j}, \widehat{a}_{1j}, \lambda_j = 0), \quad j \in \mathcal{B},$$
$$p_{ij}(\theta) = p(Y_{ij}|\theta, \widehat{d}_j, \widehat{a}_{0j}, a_{1j} = 0, \lambda_j = 0), \quad j \in [J]\backslash\mathcal{B}.$$

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

2.3. A Special Case: Linear Model with Closed-form Solution

When the link function g is the identity function, model (1) becomes the linear regression model. As DIF is defined as the group difference of the distribution of Y conditional on the latent trait, we conduct our DIF analysis in the orthogonal subspace of $\boldsymbol{\theta}$ in the linear model. To be more specific, we consider the residuals of $\mathbf{Y}, \mathbf{Z}, \mathbf{X}$ after regressing on $(1, \boldsymbol{\theta})$ respectively, denoted by $\mathbf{Y}^\dagger, \mathbf{Z}^\dagger, \mathbf{X}^\dagger$.

The model with and without the grouping variable can be rewritten as a reduced and a full linear regression model

$$\mathbf{Y}^\dagger = a_1' \boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{7}$$

$$\mathbf{Y}^\dagger = a_1 \boldsymbol{\eta} + \lambda \mathbf{Z}^\dagger + \boldsymbol{\delta}, \tag{8}$$

where $\boldsymbol{\eta} = \mathbf{X}^\dagger \boldsymbol{\omega}$. If we assume $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma_\epsilon^2)$ and $\delta_i \overset{i.i.d}{\sim} N(0, \sigma_\delta^2)$, the objective function (3) is equivalent to

$$L(\boldsymbol{\omega}) = \frac{n}{2} \log(\|\widehat{\boldsymbol{\epsilon}}\|^2) - \frac{n}{2} \log(\|\widehat{\boldsymbol{\delta}}\|^2), \tag{9}$$

where $\widehat{\boldsymbol{\epsilon}}$ and $\widehat{\boldsymbol{\delta}}$ are the linear regression residuals in (7) and (8). The zero of the objective (9) turns out to have a closed form expression under some weak conditions, in which case DIF can be fully removed. Without loss of generality, we assume that all the features are orthogonal and scaled, i.e. $\mathbf{X}^{\dagger\top}\mathbf{X}^\dagger = \mathbf{I}_K$. This is achieved by principal component analysis in practice. We also assume that $\mathbf{Y}^{\dagger\top}\mathbf{Z}^\dagger > 0$.

Proposition 1. Assume that $\mathbf{X}^{\dagger\top}\mathbf{X}^\dagger = \mathbf{I}_K$. Let $\widehat{\mathbf{Y}} = \mathbf{X}^{\dagger\top}\mathbf{Y}^\dagger$ and $\widehat{\mathbf{Z}} = \mathbf{X}^{\dagger\top}\mathbf{Z}^\dagger$. Under the condition that

$$\frac{-\|\widehat{\mathbf{Y}}\|\|\widehat{\mathbf{Z}}\| + \widehat{\mathbf{Y}}^\top\widehat{\mathbf{Z}}}{2} < \mathbf{Y}^{\dagger\top}\mathbf{Z}^\dagger < \frac{\|\widehat{\mathbf{Y}}\|\|\widehat{\mathbf{Z}}\| + \widehat{\mathbf{Y}}^\top\widehat{\mathbf{Z}}}{2}, \tag{10}$$

there exists $\widehat{\boldsymbol{\omega}}$ such that $\|\widehat{\boldsymbol{\omega}}\| = 1$ and $L(\widehat{\boldsymbol{\omega}}) = 0$. Specifically, denote

$$\alpha = \sqrt{\frac{2\mathbf{Y}^{\dagger\top}\mathbf{Z}^\dagger + \|\widehat{\mathbf{Y}}\|\|\widehat{\mathbf{Z}}\| - \widehat{\mathbf{Y}}^\top\widehat{\mathbf{Z}}}{2\|\widehat{\mathbf{Y}}\|\|\widehat{\mathbf{Z}}\|}}, \ \beta = \sqrt{\frac{-2\mathbf{Y}^{\dagger\top}\mathbf{Z}^\dagger + \|\widehat{\mathbf{Y}}\|\|\widehat{\mathbf{Z}}\| + \widehat{\mathbf{Y}}^\top\widehat{\mathbf{Z}}}{2\|\widehat{\mathbf{Y}}\|\|\widehat{\mathbf{Z}}\|}},$$

$$\boldsymbol{q}_1 = \frac{\|\widehat{\mathbf{Y}}\|\widehat{\mathbf{Z}} + \|\widehat{\mathbf{Z}}\|\widehat{\mathbf{Y}}}{\left\|\|\widehat{\mathbf{Y}}\|\widehat{\mathbf{Z}} + \|\widehat{\mathbf{Z}}\|\widehat{\mathbf{Y}}\right\|}, \ \boldsymbol{q}_2 = \frac{\|\widehat{\mathbf{Y}}\|\widehat{\mathbf{Z}} - \|\widehat{\mathbf{Z}}\|\widehat{\mathbf{Y}}}{\left\|\|\widehat{\mathbf{Y}}\|\widehat{\mathbf{Z}} - \|\widehat{\mathbf{Z}}\|\widehat{\mathbf{Y}}\right\|}.$$

Then $\widehat{\boldsymbol{\omega}} = \alpha\boldsymbol{q}_1 + \beta\boldsymbol{q}_2$ or $\widehat{\boldsymbol{\omega}} = \alpha\boldsymbol{q}_1 - \beta\boldsymbol{q}_2$ satisfies $L(\widehat{\boldsymbol{\omega}}) = 0$.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

We note that $L(\boldsymbol{\omega})$ has multiple zeros, as established in Proposition 1. We choose $\widehat{\boldsymbol{\omega}}_1 = \alpha \boldsymbol{q}_1 + \beta \boldsymbol{q}_2$ over $\widehat{\boldsymbol{\omega}}_2 = \alpha \boldsymbol{q}_1 - \beta \boldsymbol{q}_2$ for the following reason. When $\mathbf{X}^\dagger$ predicts $\mathbf{Y}^\dagger$ with high accuracy, $\mathbf{X}^\dagger \widehat{\boldsymbol{\omega}}_1$ aligns with the projection of $\mathbf{Z}^\dagger$ onto the column space of $\mathbf{X}^\dagger$, whereas $\mathbf{X}^\dagger \widehat{\boldsymbol{\omega}}_2$ aligns with the projection of $\mathbf{Y}^\dagger$ onto the same subspace. As the goal is to reduce DIF, we choose to use $\widehat{\boldsymbol{\omega}}_1$ over $\widehat{\boldsymbol{\omega}}_2$. For further details, see the proof of Proposition 1 in the Appendix.

There are several scenarios in which condition (10) holds. The first scenario occurs when $\mathbf{Y}^{\dagger\top}\mathbf{Z}^\dagger = 0$, which arises if the response and the grouping variable are independent conditional on the target trait, indicating that the item is not a DIF item, and condition (10) is automatically satisfied. The second scenario is when $\mathbf{X}^\dagger$ has high linear predictability of $\mathbf{Y}^\dagger$. Specifically, when $\mathbf{X}^\dagger$ predicts $\mathbf{Y}^\dagger$ with high accuracy, we find that $\mathbf{X}^\dagger \mathbf{X}^{\dagger\top} \mathbf{Y}^\dagger \approx \mathbf{Y}^\dagger$, which leads to $\widehat{\mathbf{Y}}^\top \widehat{\mathbf{Z}} = \mathbf{Y}^{\dagger\top} \mathbf{X}^\dagger \mathbf{X}^{\dagger\top} \mathbf{Z}^\dagger \approx \mathbf{Y}^{\dagger\top} \mathbf{Z}^\dagger$, making it straightforward to verify condition (10). This is the case where process data capture the nuisance factors that affect the response. In the simplest case, we assume $\mathbf{X}$ can linearly predict $\mathbf{Y}$ with full accuracy and let $\mathbf{Y} = \mathbf{X}\mathbf{A}$ with some vector $\mathbf{A} \in \mathbb{R}^K$. Then,

$$\mathbf{Y}^\dagger = \mathbf{Y} - \mathbb{E}[\mathbf{Y}|\boldsymbol{\theta}] = (\mathbf{X} - \mathbb{E}[\mathbf{X}|\boldsymbol{\theta}])\,\mathbf{A} = \mathbf{X}^\dagger \mathbf{A}.$$

Accordingly, we consider this case to be achievable. The third scenario is when $\mathbf{X}$ has high linear predictability of $\mathbf{Z}$. Following similar calculations, we conclude that condition (10) is satisfied. With that being said, it is generally difficult to predict $\mathbf{Z}$ with process data.

2.4. General Cases

The main focus of this paper is to reduce DIF of univariate Z. Nonetheless, we have a brief extension to some simple cases of multi-dimensional Z. While we have previously focused on uniform DIF in a linear factor model with one grouping variable, the proposed method is applicable to other general cases. In what follows, we will shift our focus to addressing non-uniform DIF, continuous covariates, multiple grouping variables , and nonlinear models. The goal is to minimize the objective function (3).

Non-uniform DIF. Non-uniform DIF occurs when not only the intercept parameter, but also the discrimination parameter (the coefficient of θ_i) differs across groups. More specifically, for

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

non-uniform DIF, (1) becomes

$$g(\mu_i) = d + a_0\theta_i + a_1\eta_i + \lambda Z_i + \lambda' Z_i\theta_i. \tag{11}$$

Non-uniform DIF can be viewed as a special case involving one grouping variable Z_i and a continuous covariate $Z_i\theta_i$. Therefore, non-uniform DIF is included in the continuous covariates and multiple-group cases discussed below.

Continuous covariates. In some applications, DIF is brought by continuous covariates. For instance, in computer-based tests, age is a very important variable. As our proposed method does not require Z_i to be a discrete variable, it is applicable when Z_i is a continuous variable. To see this, let $Z_i \in \mathbb{R}$, and we aim to construct $\widehat{\eta}_{ij}$ such that

$$p(Y_i = y|\theta_i, \widehat{\eta}_i, Z_i) \approx p(Y_i = y|\theta_i, \widehat{\eta}_i). \tag{12}$$

When (12) holds, we expect the objective function (3) to be close to 0. Therefore, minimizing (3) to reduce DIF is valid for continuous covariates.

Nonlinear models. When the link function $g(\cdot)$ is not linear, e.g. the M2PL model and the Probit regression model, minimizing the objective function (3) is a nested optimization problem that takes in different forms depending on the model employed. Again, we rely on numerical methods to approximate the solution.

Multiple grouping variables. Sometimes it is of interest to evaluating DIF over more than one grouping variables (Kim et al., 1995; Bauer et al., 2020). For cases involving multiple grouping variables, the expression of the objective function in terms of $\boldsymbol{\omega}$ becomes significantly more complex. To address this, we propose an approximation of the objective function for M grouping variables $\mathbf{Z}_1, \ldots, \mathbf{Z}_M \in \mathbb{R}^N, M \geq 2$ as follows:

$$L(\boldsymbol{\omega}) = \sum_{m=1}^{M} L^{(m)}(\boldsymbol{\omega}), \tag{13}$$

where $L^{(m)}(\boldsymbol{\omega})$ is the objective function (3) corresponding to $\mathbf{Z}_m$. In general, there is no closed-form solution for the minimizer of (3) or (13) with the presence of multiple grouping variables. Therefore, we rely on numerical methods to approximate the minimizer.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

2.5. Procedure

We outline the procedure of the proposed method in a practical setting with J items.

1. Suppose we have access to a set of anchor items that are DIF-free, which are used to perform DIF detection on all the items. Suppose DIF has been detected on a subset of items $\mathcal{B} \subset [J]$.
2. Obtain an initial latent trait estimate $\widehat{\boldsymbol{\theta}}^{(0)}$ using the items without DIF.
3. For each item, perform the proposed method to obtain the nuisance trait surrogates and DIF-corrected model parameters. More specifically, for each item $j \in \mathcal{B}$,
   (a) With the initial ability estimate, $\widehat{\boldsymbol{\theta}}^{(0)}$, find the minimizer $\widehat{\boldsymbol{\omega}}_j$ in Equation (5) and obtain the nuisance trait estimate $\widehat{\boldsymbol{\eta}}_j = \mathbf{X}_j\widehat{\boldsymbol{\omega}}_j$.
   (b) With $\widehat{\boldsymbol{\theta}}^{(0)}, \widehat{\boldsymbol{\eta}}_j$, obtain the item parameter estimates $\widehat{d}_j, \widehat{a}_{0j}, \widehat{a}_{1j}$ in (1) with λ_j set to 0, using the full data.
4. Obtain the updated estimate of θ_i with (6), utilizing the nuisance trait surrogates and the calibrated measurement models, denoted by $\widehat{\boldsymbol{\theta}}^{(1)}$.

The above procedure requires an initial estimate $\widehat{\boldsymbol{\theta}}^{(0)}$ that does not have DIF. In case that DIF-free items are not available, we could apply the above method iteratively. In particular, if $\widehat{\boldsymbol{\theta}}^{(0)}$ is known to have some DIF, we apply the above procedure (steps 1 through 4) and obtain $\widehat{\boldsymbol{\theta}}^{(1)}$. Generally speakly, $\widehat{\boldsymbol{\theta}}^{(1)}$ has less DIF than $\widehat{\boldsymbol{\theta}}^{(0)}$. We set the updated estimate $\widehat{\boldsymbol{\theta}}^{(1)}$ the initial value and apply steps 1 through 4 and obtain $\widehat{\boldsymbol{\theta}}^{(2)}$. We could iteratively apply the procedure until some type of convergence is reached. The convergence could be measured by some difference (or similarity) metric between two iterations: $\widehat{\boldsymbol{\theta}}^{(n)}$ and $\widehat{\boldsymbol{\theta}}^{(n+1)}$. In our real data example, we used the sample correlation and two iterations is sufficient. The correlation between $\widehat{\boldsymbol{\theta}}^{(1)}$ and $\widehat{\boldsymbol{\theta}}^{(2)}$ is over 99%. Researchers may use other metrics depending on the specific scope of their study, such as, average difference, maximum difference, and general L_p distance, etc.

3. Simulation Studies

We carry out extensive simulation experiments to evaluate the proposed method in this section. The goal is to show that the proposed method is able to minimize the objective function,

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

accurately estimate the nuisance traits and the item parameters, and correct for target trait estimation from the DIF items.

3.1. Simulation Settings

We consider three settings with the sample sizes $N = 2000, 5000, 10000$. Among the subjects, $2/3$ are in the reference group and $1/3$ are in the focal group. We fix the number of items as $J = 15$ and consider low, medium, high proportions of DIF items, that is, $3, 5, 10$ DIF items. We also consider two settings for the DIF parameters a_{1j}. For small DIF effects, a_{1j}s are uniformly sampled from 0.5 to 1; for large DIF effects, the range is from 1 to 1.5. In summary, there are 18 simulation settings varying in sample size, proportion of DIF items, and DIF effect size. For each simulation setting, 100 independent replications are generated. In each replication, we independently sample the target latent trait θ_i from $N(-0.5, 1)$ for the focal group, and $N(0.5, 1)$ for the reference group. The difficulty parameters d_j are sampled uniformly from -1 to 1 and the discrimination parameters a_{0j} are sampled uniformly from 1 to 2. As the generation of process data is challenging, we generate the process data features directly. The number of process data features for each item is fixed at $K = 100$. The process data features $\mathbf{x}_{ij} \in \mathbb{R}^K$ are first independently sampled from the multivariate Gaussian distribution $N(\boldsymbol{\mu}^{(Z_i)}, \mathbf{I}_K)$. The mean of the process features depends on the respondent's group membership. For the reference group, $\boldsymbol{\mu}^{(0)} = \mathbf{1}_K$, and for the focal group $\boldsymbol{\mu}^{(1)} = -\mathbf{1}_K$. We then right multiply $\mathbf{X}_j$ with sample $\mathrm{Cov}(\mathbf{X}_j)^{-1/2}$ for the process features to have the identity matrix as the covariance. The ground truth nuisance trait is a linear combination of the process features: $\eta_{ij} = \boldsymbol{\omega}_j^\top \mathbf{x}_{ij}$, with ω_{jk} sampled independently from the exponential distribution with rate 1 and $\boldsymbol{\omega}_j$'s are scaled to have unit norm. Note that the generated nuisance traits for each item have unit norm. We consider both the linear model and the M2PL model in generating the item responses. As DIF is solely introduced by the distributional difference of the nuisance trait in simulation, λ_j is set to be 0 in (1) when generating the item responses for both models.

For the linear factor model, we set the variance of the item response noise as 1 when generating the item responses. To make sure that the features can almost perfectly predict each item response in the linear model, we add one column of $\mathbf{Y}_j + N(0, 0.1 \cdot \mathbf{I}_n)$ to the generated $\mathbf{X}_j$

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

for each item j and generate the nuisance traits with the same manner as mentioned above. The initial target trait estimates $\widehat{\theta}_i^{(0)}$ are obtained with factor analysis using responses from the DIF-free items. To construct the nuisance trait, we adopt the closed form expression of $\widehat{\boldsymbol{\omega}}$ from Proposition 1. For the M2PL model, we generate Y_{ij} with the link function $g(\cdot)$ being the the logit function. The initial target trait estimates $\widehat{\theta}_i^{(0)}$ are the MLE estimates using the DIF-free items. The nuisance traits are estimated by solving (5) using the `optim` function with the L-BFGS-B optimizer in `R`.

The simulation above corresponds to uniform DIF. In addition, we consider the case with non-uniform DIF so that λ' in (11) is not zero. Similar simulation settings are adopted except for the generation of process data features. To ensure the existence of non-uniform DIF, the process features $\mathbf{x}_{ij}$ are sampled from the multivariate Gaussian distribution $N(\boldsymbol{\mu}^{(Z_i)} + \gamma_j \theta_i Z_i, \mathbf{I}_K)$ with γ_j simulated from the exponential distribution with rate 1.

3.2. Evaluation Criteria

We consider five evaluation criteria. Firstly, we check whether the proposed method is able to reduce the objective function value. Specifically, we verify whether zero of (3) is obtained for the linear model with Proposition 1. Secondly, we evaluate the correlation of nuisance trait estimation compared to its ground truth. Thirdly, we calculate the mean squared error (MSE) of the item parameter estimations. In addition, to evaluate measurement reliability after changing the scoring rule, we calculate the Fisher information (FI) of the target trait for the DIF items. For the linear model, target trait FI for the item $j \in \mathcal{B}$ is $\text{FI}_j = \widehat{a}_{0j}^2 / \widehat{\sigma}_j^2$, where $\widehat{a}_{0j}$ is the estimated coefficient for the target trait and $\widehat{\sigma}_j^2$ is the estimated variance. For the M2PL model, $\text{FI}_j = \frac{1}{N} \sum_{i=1}^N \widehat{p}_{ij}(1 - \widehat{p}_{ij}) \widehat{a}_{0j}^2$, where $\widehat{p}_{ij}$ is the estimated response of the logistic model. Last but not least, we consider the between-group sum of squared (SS) bias for target trait estimation using the DIF items. We elaborate more on this evaluation criterion assuming one grouping variable with a focal group and a reference group. Note that this criterion can be easily generalized to multiple grouping variables. For any target trait estimate $\widetilde{\boldsymbol{\theta}}$, define the bias $\boldsymbol{\nu} := \widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}$ and denote the mean of bias as $\bar{\nu} := \frac{1}{N} \sum_{i=1}^N \nu_i$. Consider the between-group SS for

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

the bias of $\widetilde{\boldsymbol{\theta}}$ in analysis of variance (ANOVA):

$$SSB_{\widetilde{\boldsymbol{\theta}}} = N_r \left(\frac{1}{N_r} \sum_{i \in \mathcal{N}_r} \nu_i - \bar{\nu} \right)^2 + N_f \left(\frac{1}{N_f} \sum_{i \in \mathcal{N}_f} \nu_i - \bar{\nu} \right)^2, \tag{14}$$

where $\mathcal{N}_r$ and $\mathcal{N}_f$ are the index sets of the reference group and focal group respectively, and $N_r = |\mathcal{N}_r|, N_f = |\mathcal{N}_f|$. To illustrate that the proposed method is able to de-bias target trait estimation, we compare the above-defined value for two estimates. The benchmark estimate is the MLE computed using the responses from the DIF items in $\mathcal{B}$, assuming DIF is not present:

$$\check{\theta}_i = \underset{\theta}{\operatorname{argmax}} \sum_{j \in \mathcal{B}} \log p(Y_{ij} | \theta, \check{d}_j, \check{a}_{0j}, a_{1j} = 0, \lambda_j = 0),$$

where $\check{d}_j, \check{a}_{0j}$ are calibrated on the DIF items. The second estimate is the MLE computed after DIF-correction:

$$\widehat{\theta}_i = \underset{\theta}{\operatorname{argmax}} \sum_{j \in \mathcal{B}} \log p(Y_{ij} | \theta, \widehat{d}_j, \widehat{a}_{0j}, \widehat{a}_{1j}, \lambda_j = 0),$$

where $\widehat{d}_j, \widehat{a}_{0j}, \widehat{a}_{1j}$ are calibrated on the DIF items with the nuisance trait surrogates. Because of the presence of DIF, $\check{\boldsymbol{\theta}}$ is expected to be over-estimated within one group, and under-estimated within the other, leading to large values of $SSB_{\check{\boldsymbol{\theta}}}$. With the proposed method, we expect $SSB_{\widehat{\boldsymbol{\theta}}}$ for the corrected estimate to be small compared to $SSB_{\check{\boldsymbol{\theta}}}$ for the not-corrected estimate.

3.3. Simulation Results

Figure 2 summarizes the values for the objective function (3) before and after adding the nuisance trait surrogate for the linear and M2PL models, with uniform DIF, $N = 500$, and large DIF. For complete results of all simulation settings, see Figures 8 and 9 in the Appendix. We are able to minimize the objective function across simulation settings and replications. Specifically, we are able to obtain the zero of the objective function for the linear model. Tables 3.3 and 3.3 demonstrate the MSE of the item parameter estimates and the correlation between the ground truth and estimated nuisance traits. We observe that the MSE values are small in magnitude and the nuisance trait estimation correlation is high. It also shows that as the sample size increases, MSE tends to decrease and the nuisance trait correlation tends to increase. On the other hand,

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

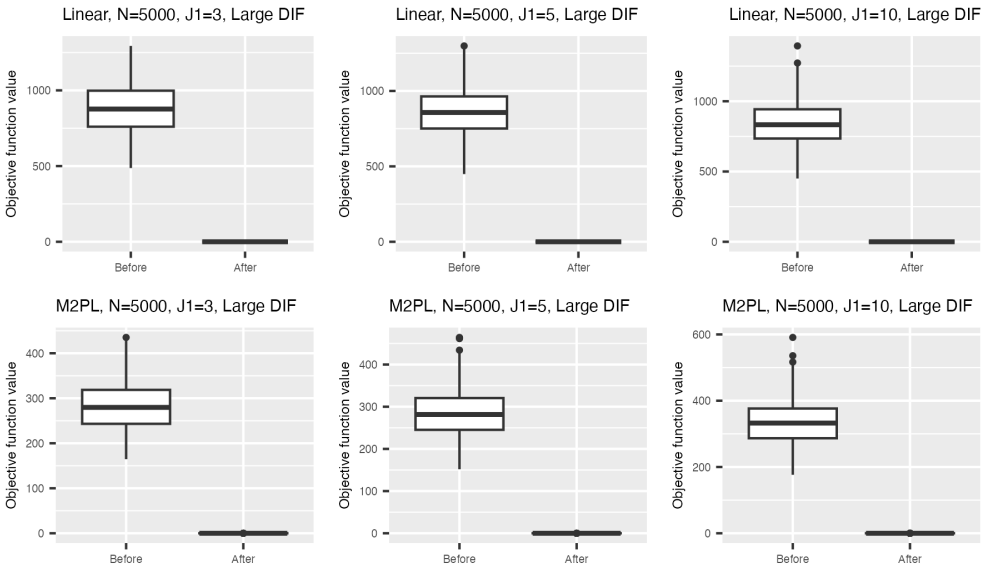

FIGURE 2.

Values of the objective function before and after adding the nuisance trait surrogate for the linear model (upper) and the M2PL model (lower) with *uniform* DIF, $N = 5000$, and large DIF.

larger DIF effects and DIF item proportion lead to larger MSE. For the linear model, the nuisance trait correlation does not change with the sample size. For the M2PL model, it increases with the sample size. Figures 10 and 11 in the Appendix summarize the Fisher information of the target trait θ before and after adding the nuisance trait surrogate for the linear and M2PL models respectively. We observe an increase in Fisher information of the target trait in the measurement model after correcting for DIF in both models. Furthermore, Figure 3 compares the between-group SS bias of the corrected and not-corrected target trait estimates using the DIF items, for the linear and M2PL models. The x-axis corresponds to the estimation without DIF correction, and the y-axis corresponds to the DIF-corrected estimation. The corresponding simulation setting is $N = 500$ with large DIF; for the complete results, see Figures 12 and 13 in the Appendix. We see that the proposed method is able to correct for the target estimation bias introduced by DIF, as the between-group SS bias is significantly reduced after DIF correction for both models.

Results for non-uniform DIF are deferred to the Appendix. In Figures 14, 15, we observe an

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

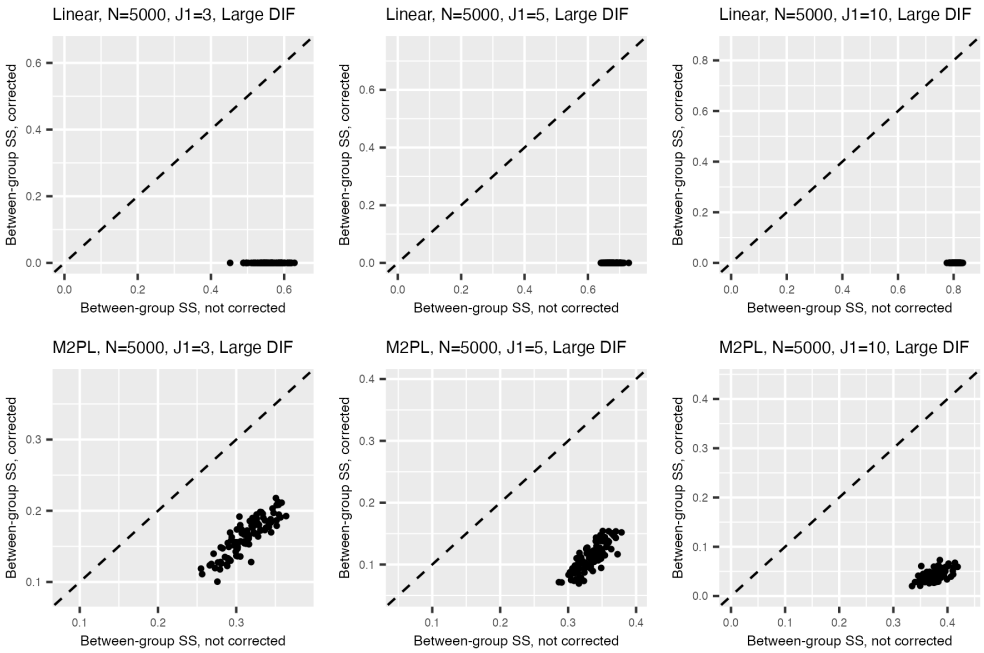

FIGURE 3.

Between-group sum of squared bias for target trait estimation for the linear model (upper) and the M2PL model (lower) with *uniform* DIF, $N = 5000$, and large DIF. The x-axis corresponds to the estimation without DIF correction using the DIF items; the y-axis corresponds to the DIF-corrected estimation using the DIF items.

increase in the minimized objective function values for non-uniform compared to uniform DIF, although the proposed method is still able to reduce DIF significantly. Similar results for item parameter and nuisance trait estimation, target trait Fisher information, and between-group SS bias are observed; see Tables 8, 8, Figures 16, 17, 18, 19 in the Appendix.

4. Case Study

We use the PIAAC 2012 survey data as a case study to demonstrate the effectiveness of our method. Response processes of 13 Problem Solving in Technology-Rich Environments (PSTRE) items from 17 countries are considered in this study. PIAAC was the first attempt to assess the PSTRE construct on a large scale and as a single dimension. Under the PIAAC framework, PSTRE is defined as the use of digital technology, communication tools, and the internet to obtain and evaluate information, communicate with others, and perform practical tasks (OECD,

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

		Small DIF			Large DIF		
		J_1=3	J_1=5	J_1=10	J_1=3	J_1=5	J_1=10
MSE (d)	$N = 2000$	0.0365	0.0387	0.0363	0.1037	0.1055	0.0997
	$N = 5000$	0.0377	0.0387	0.0375	0.1042	0.0969	0.1008
	$N = 10000$	0.0375	0.0375	0.0377	0.1005	0.0999	0.0990
MSE (a_0)	$N = 2000$	0.0014	0.0013	0.0014	0.0017	0.0020	0.0019
	$N = 5000$	0.0007	0.0005	0.0006	0.0007	0.0007	0.0009
	$N = 10000$	0.0003	0.0003	0.0004	0.0003	0.0004	0.0004
MSE (a_1)	$N = 2000$	0.0494	0.0502	0.0495	0.1292	0.1344	0.1354
	$N = 5000$	0.0497	0.0492	0.0493	0.1372	0.1335	0.1346
	$N = 10000$	0.0489	0.0494	0.0498	0.1335	0.1350	0.1324
Corr ($\widehat{\boldsymbol{\eta}}, \boldsymbol{\eta}$)	$N = 2000$	0.7112	0.7116	0.7125	0.7150	0.7103	0.7112
	$N = 5000$	0.7115	0.7113	0.7140	0.7107	0.7107	0.7115
	$N = 10000$	0.7162	0.7127	0.7128	0.7134	0.7110	0.7127

TABLE 1.

Mean squared error of item parameter estimates and nuisance trait correlation for the *linear model* with *uniform* DIF under different simulation settings. The values are averaged across the DIF items and replications.

2012b). The survey also recorded a broad spectrum of respondents' background information such as gender, age, occupation, hourly income, education level, etc.. To conduct DIF analysis, we consider age, income, and gender as the demographic grouping variables. We include the process data of 8,398 respondents who answered all 13 items and have no missing value of the three covariates in the study. For age, we use 47, which is the 70% quantile, as the cutoff value to split the samples into younger and older sub-populations. The younger population is treated as the reference group, and the older population is treated as the focal group. For income, we first group the samples by their nationality, and then use the medium income of each nation as the cutoff value. The lower-income and higher-income groups are treated as the focal and the reference groups, respectively. For gender, female is treated as the focal group and male as the reference group.

Table 3 provides a descriptive summary of the 13 items, where n is the number of total possible actions, $\bar{L}$ is the average process sequence length, and Correct % is the percentage

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

		Small DIF			Large DIF		
		J_1=3	J_1=5	J_1=10	J_1=3	J_1=5	J_1=10
MSE (d)	$N = 2000$	0.0365	0.0387	0.0363	0.1037	0.1055	0.0997
	$N = 5000$	0.0377	0.0387	0.0375	0.1042	0.0969	0.1008
	$N = 10000$	0.0375	0.0375	0.0377	0.1005	0.0999	0.0990
MSE (a_0)	$N = 2000$	0.0014	0.0013	0.0014	0.0017	0.0020	0.0019
	$N = 5000$	0.0007	0.0005	0.0006	0.0007	0.0007	0.0009
	$N = 10000$	0.0003	0.0003	0.0004	0.0003	0.0004	0.0004
MSE (a_1)	$N = 2000$	0.0494	0.0502	0.0495	0.1292	0.1344	0.1354
	$N = 5000$	0.0497	0.0492	0.0493	0.1372	0.1335	0.1346
	$N = 10000$	0.0489	0.0494	0.0498	0.1335	0.1350	0.1324
Corr ($\widehat{\boldsymbol{\eta}}, \boldsymbol{\eta}$)	$N = 2000$	0.7112	0.7116	0.7125	0.7150	0.7103	0.7112
	$N = 5000$	0.7115	0.7113	0.7140	0.7107	0.7107	0.7115
	$N = 10000$	0.7162	0.7127	0.7128	0.7134	0.7110	0.7127

TABLE 2.

Mean squared error of item parameter estimates and nuisance trait correlation for the *linear model* with *uniform* DIF under different simulation settings. The values are averaged across the DIF items and replications.

receiving the full credit on each item. When solving for each item, the respondents are presented with one or more simulated informational and communicative (ICT) environments, such as an email inbox, a spreadsheet, a web browser, etc. For example, in item U01a, the respondents are presented with an email inbox interface and are asked to classify the email senders into 'can come' and 'cannot come' categories based on their email contents. To complete the task, the respondents need to conduct a sequence of clicking, dragging, or typing actions, which are recorded in the log files as process data. In addition, Tables 8 and 8 in the Appendix summarize the mean and standard deviation of the responses by group for each item, for the polytomous responses and binary responses respectively.

We use the MDS approach proposed in Tang et al. (2020a) to extract item features that approximate the geometric distances defined by a dissimilarity matrix of the action sequences. We set the dimension of features to be $K = 100$ to ensure enough information is retained in the extracted features. To verify that the features contain an adequate amount of information in the

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

Item ID	Description	n	$\bar{L}$	Correct %
U01a	Party Invitations – Can/Cannot Come	51	17.2	59.8
U01b	Party Invitations – Accommodations	55	26.1	52.4
U02	Meeting Rooms	100	26.9	15.7
U03a	CD Tally	67	9.0	42.3
U04a	Class Attendance	926	39.2	15.3
U06a	Sprained Ankle – Site Evaluation Table	30	9.5	26.2
U06b	Sprained Ankle – Reliable/Trustworthy Site	26	15.0	50.8
U07	Digital Photography Book Purcha	40	18.6	51.7
U11b	Locate E-mail – File 3 E-mails	137	24.8	26.5
U16	Reply All	886	32.4	63.9
U19a	Club Membership – Member ID	162	17.3	75.1
U19b	Club Membership – Eligibility for Club President	450	20.7	52.5
U23	Lamp Return	164	21.7	38.2

TABLE 3.

Summary statistics of 13 PIAAC problem-solving items. Here n is the number of total possible actions, $\bar{L}$ is the average process sequence length, and Correct % is the percentage of correct answers.

process data, we use them to predict the responses with both linear regression and logistic regression. Results show that the extracted features can perfectly predict whether the examine received the full credit of each item with both regression methods. Engagement is often considered a nuisance trait in respondent behavior (Wise and Kong, 2005; Wise and DeMars, 2006), and one plausible measure of engagement is the length of a respondent's process sequence. For each item, we randomly sample 80% of training data to predict the process sequence length with the extracted features using ridge regression, where the ridge parameter is selected with cross-validation on the training data. Figure 4 demonstrates the out-of-sample correlation

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

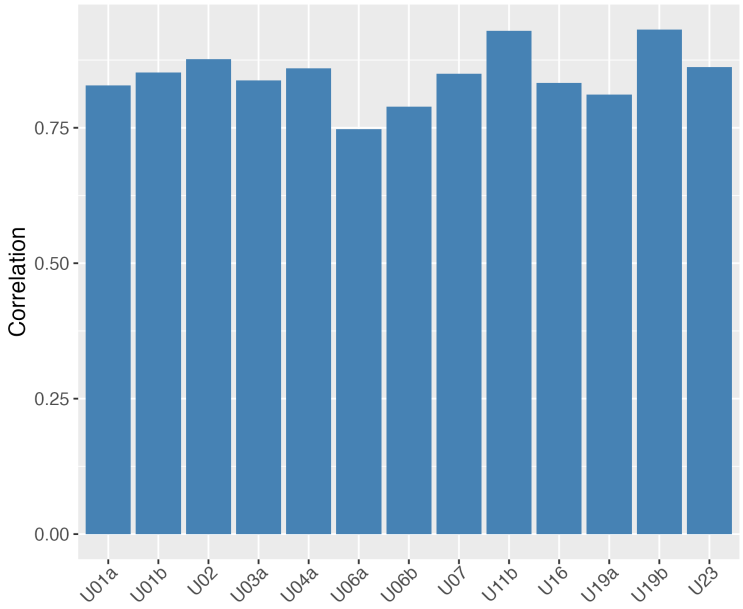

FIGURE 4.

Out-of-sample prediction correlation of the process sequence length using the extracted features for each item.

between the predicted and actual process sequence lengths for each item. It shows that the extracted features are able to predict the sequence length with high accuracy.

4.1. DIF Existence

The proposed procedure assumes that we have access to the ground truth target trait values. However, these target traits are unknown and need to be estimated in practice. To obtain initial estimates of the target traits, we utilize the responses to all 13 items. We then use these target trait estimates to identify DIF items. For the linear model, the latent traits are estimated by performing maximum-likelihood factor analysis on the response data. For the M2PL model, the link function $g(\cdot)$ in Equation (1) is the logit function, and the reduced model without the nuisance trait is calibrated by maximizing the marginal likelihood using the expectation-maximization algorithm (Bock and Aitkin, 1981). The initial estimation $\widehat{\boldsymbol{\theta}}^{(0)}$ is the MLE estimate after item calibration.

Table 4 summarizes the DIF detection results using the initial trait estimate $\widehat{\boldsymbol{\theta}}^{(0)}$. For the age

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

variable, 12 out of 13 items are detected to have uniform DIF with both the linear model and the logistic model. For the income variable, 5 items have uniform DIF with both models. For the gender variable, 5 are uniform-DIF items with the linear model and 7 with the logistic model.

Item	$\widehat{\lambda}$ ($\widehat{\sigma}(\widehat{\lambda})$)					
	Linear model			Logistic model		
	Age	Income	Gender	Age	Income	Gender
U01a	**-0.256(0.018)**	**0.032(0.016)**	**-0.040(0.016)**	**-0.861(0.067)**	0.117(0.064)	**-0.147(0.063)**
U01b	**-0.074(0.018)**	0.012(0.017)	-0.005(0.016)	**-0.238(0.065)**	0.023(0.060)	0.003(0.059)
U02	**0.088(0.021)**	**-0.041(0.019)**	-0.019(0.019)	**0.377(0.092)**	**-0.152(0.076)**	0.022(0.076)
U03a	**-0.076(0.020)**	-0.020(0.018)	**-0.093(0.018)**	**-0.231(0.063)**	-0.067(0.056)	**-0.266(0.056)**
U04a	**0.089(0.022)**	**-0.05(0.020)**	-0.032(0.020)	**0.345(0.082)**	**-0.161(0.069)**	-0.054(0.069)
U06a	**0.123(0.021)**	**-0.045(0.019)**	**0.039(0.019)**	**0.421(0.068)**	**-0.133(0.059)**	**0.172(0.059)**
U06b	**0.063(0.023)**	-0.037(0.021)	0.028(0.021)	**0.164(0.052)**	-0.082(0.047)	0.067(0.046)
U07	**0.132(0.020)**	-0.023(0.018)	**0.076(0.018)**	**0.426(0.061)**	-0.080(0.054)	**0.239(0.054)**
U11b	**-0.051(0.021)**	0.023(0.019)	0.026(0.019)	**-0.207(0.070)**	0.081(0.059)	**0.134(0.059)**
U16	**-0.037(0.019)**	**0.053(0.017)**	**0.052(0.017)**	-0.078(0.065)	**0.164(0.061)**	**0.176(0.061)**
U19a	**0.069(0.020)**	-0.001(0.018)	**0.046(0.018)**	**0.347(0.070)**	-0.025(0.066)	0.127(0.065)
U19b	**0.074(0.018)**	0.012(0.017)	**-0.076(0.016)**	**0.269(0.065)**	0.016(0.059)	**-0.257(0.059)**
U23	-0.033(0.020)	**0.051(0.018)**	-0.001(0.018)	**-0.115(0.064)**	**0.157(0.056)**	0.034(0.056)

TABLE 4.

DIF detection results *without* the nuisance trait. Bold text indicates statistical significance under the 0.05 significance level.

4.2. DIF Correction

We implement the proposed method to estimate the nuisance traits with the extracted process features, and to correct for DIF effects. For the linear model, we use the closed-form expression in Proposition 1 as the estimate, while for the M2PL model, the nuisance traits are estimated by solving (5) using the `optim` function with the L-BFGS-B optimizer in `R`. We also apply the iterative method in Section 2.5, the correlations between $\widehat{\boldsymbol{\theta}}^{(1)}$ and $\widehat{\boldsymbol{\theta}}^{(2)}$ are over 99%.

Figure 5 includes the boxplots of the objective function (3) with and without the nuisance

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

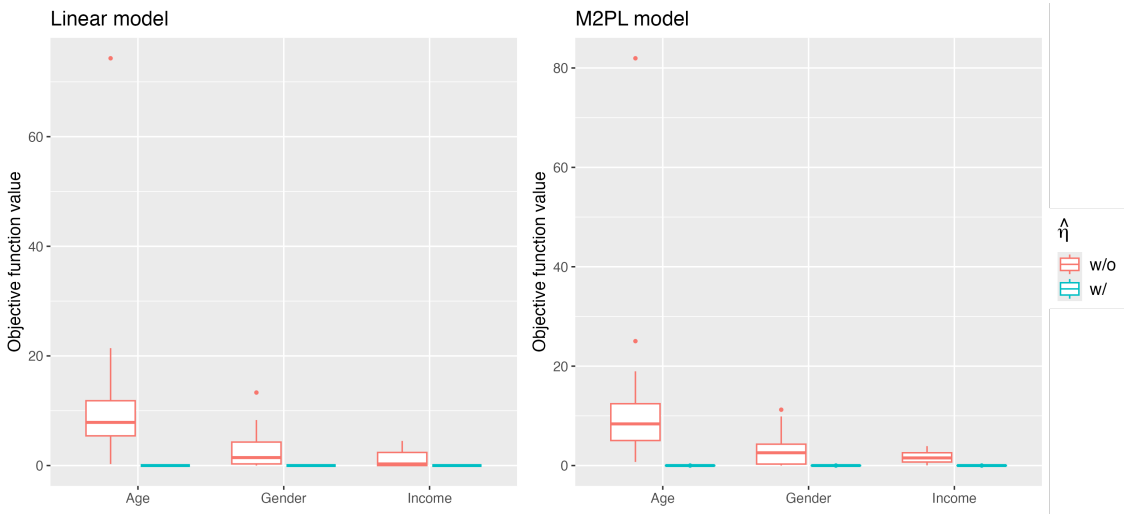

FIGURE 5.

Comparing the objective function value with and without the nuisance trait surrogate for the linear model (left) and the M2PL model (right) with one grouping variable.

trait surrogate for the three grouping variables. We see that the estimated nuisance traits serve the purpose of minimizing the objective functions for both models. Table 5 demonstrates the sample mean Fisher information of the target trait θ before and after adding the nuisance trait surrogate $\widehat{\boldsymbol{\eta}}$ for items that exhibit DIF. We see that the Fisher information increases in the linear model by adding the nuisance trait surrogate. For the M2PL model, we observe moderate reduction for most items, and significant reduction for items U06b and U07. For these two items, adding the nuisance trait surrogate significantly increases the the prediction accuracy of the item response. The boxplots of the objective function with and without the nuisance trait when two grouping variables are present can be found in Figure 20 in the Appendix. We see that the estimated nuisance traits are able to minimize the objective function with two grouping variables. For the case study, we do not have access to the ground truth target trait or a reliable unbiased estimator of the target trait as many items exhibit DIF. Therefore, we do not compare the between-group SS bias for the target trait estimates before and after DIF correction.

As an illustration, we consider item U01a and the M2PL model to interpret the results for the age variable. After obtaining the estimated nuisance traits, we update the estimation of the

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

Item ID	Linear model				M2PL model			
	w.o. $\widehat{\eta}$	Age	Income	Gender	w.o. $\widehat{\eta}$	Age	Income	Gender
U01a	1.265	1.739	1.274	1.267	0.834	0.195	–	0.569
U01b	0.896	–	–	0.897	0.796	0.335	–	–
U02	1.183	1.231	1.294	–	0.756	0.729	0.462	–
U03a	0.623	–	0.627	0.684	0.667	0.168	–	0.176
U04a	0.312	0.337	0.322	–	0.395	0.193	0.183	–
U06a	0.345	0.397	0.361	0.351	0.489	0.146	0.285	0.248
U06b	0.133	0.137	–	–	0.146	0.030	–	0.022
U07	0.545	0.707	–	0.610	0.507	0.007	–	0.027
U11b	0.587	0.605	–	–	0.536	0.443	–	0.273
U16	0.822	0.827	0.851	0.840	0.645	–	0.512	0.292
U19a	0.573	0.587	–	0.584	0.507	0.340	–	0.426
U19b	1.520	–	1.523	–	0.750	0.143	0.625	–
U23	0.867	0.870	0.912	0.867	0.619	0.540	0.358	0.312

TABLE 5.

Sample mean Fisher information for θ with and without nuisance trait surrogate, for three grouping variables and two models. Only the values corresponding to the DIF items are present.

target trait as $\widehat{\boldsymbol{\theta}}$ by solving Eq (6). Recall that DIF arises when the functioning of the response differ given the latent trait. Therefore, we study the characteristics of the residual nuisance trait given the target trait, i.e., $\widetilde{\boldsymbol{\eta}} = \widehat{\boldsymbol{\eta}} - \mathbb{E}[\widehat{\boldsymbol{\eta}}|\boldsymbol{\theta}]$. To interpret why DIF occurs in item U01a, we check the original process sequences corresponding to the minimum and maximum values of $\widetilde{\boldsymbol{\eta}}$, and find that the usage of using drag/drop actions is related to the value of the residual nuisance trait. To verify this assumption, we calculate the correlation between $\widetilde{\boldsymbol{\eta}}$ and whether drag/drop actions are used, which achieves 0.61. And the correlation between $\widetilde{\boldsymbol{\eta}}$ and the number of drag/drop actions achieves 0.49. These results suggest that the estimated nuisance variable can indicate the intensity of using drag/drop actions. Furthermore, the item response accuracy is 74.4% among the group that used drag/drop actions, versus 26.3% among those that did not use drag/drop actions. Figure 6 demonstrates the density plots of the residual nuisance trait $\widetilde{\boldsymbol{\eta}}$ among the 'older'/'younger' groups, and among the groups that did or did not use drag/drop actions. A

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

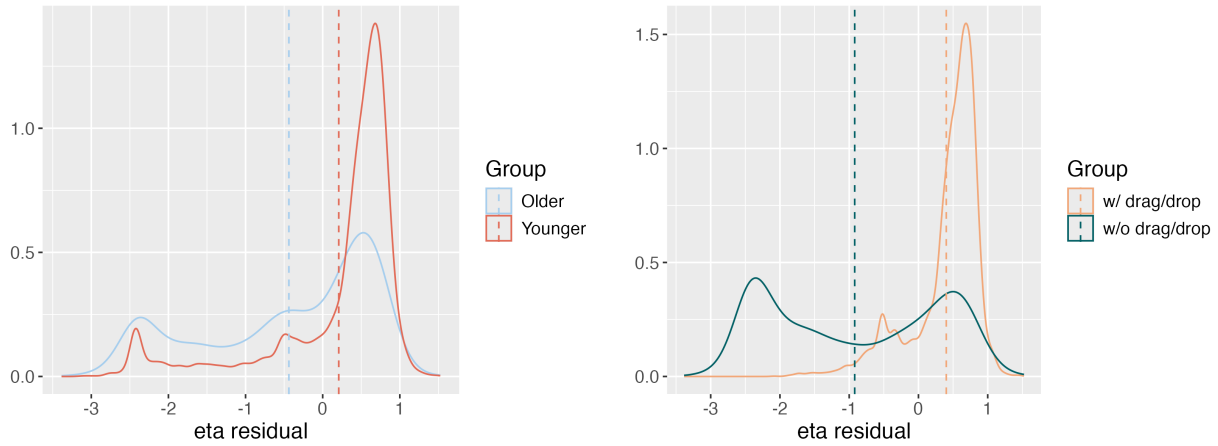

FIGURE 6.

On the left: density plot of the residual nuisance trait $\widetilde{\boldsymbol{\eta}}$ among the 'old' group and the 'young' group. On the right: density plot of the residual nuisance trait $\widetilde{\boldsymbol{\eta}}$ among the group that used drag/drop actions and those that did not.

possible interpretation to this phenomenon is that, more senior individuals might be less familiar with this type of drag-and-drop mouse usage. It is also possibly more error-prone for more senior individuals to move emails using drag-and-drop actions because of the small font size in the email interface and the narrow distances between email folders. Among the individuals who used drag/drop actions, 80.6% of the younger population had correct responses, while the percentage for the older population is only 57.7%.

Notice that majority of the items has some level of DIF. We apply the DIF correction method iteratively and found that the final estimates of θ between iteration 1 and iteration 2 is over 99%. To keep the illustration simple, we only report the results of iteration 1.

5. Discussion

Test fairness is a prominent concern within psychometric and educational research, and DIF analysis is a commonly practiced approach to ensure test fairness. When the distributions of the item response differ among two or more groups conditional on the target trait(s), DIF arises. Development of high-quality operational test items is costly, yet in practice, items identified with significant DIF effects are often discarded. In this paper, we propose a method to "de-bias" items

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

that are detected with DIF by incorporating data beyond item responses. We utilize the rich information contained in item process data, which capture the whole response processes of respondents when they interact with computer-based items. Specifically, we attribute DIF to multidimensionality, where nuisance traits with heterogeneous sub-group distributions also affect the item responses, besides the target trait to be measured. To uncover the unobserved nuisance trait, we propose to minimize the maximum likelihood difference of the models with and without the grouping variable. In the simple case with linear regression models and one grouping variable, there is a closed-form solution to the proposed optimization problem. Simulation studies and a real data case study demonstrate the effectiveness of the proposed method.

Some limitations do exist in the current method. Firstly, the assumption that DIF stems from multidimensionality may not hold, as DIF may arise from other forms of model misspecification (Huang et al., 2024). Our study focuses on how process data help capture additional latent traits contributing to DIF within the multidimensionality framework. The amount of DIF reduction is limited to the portion that is caused by nuisance attributes and their predictability by the process data. Secondly, introducing the nuisance trait into the model might reduce measurement reliability, as suggested in some decrease in the Fisher information for the target variable in the case study. It is of interest to study if a weighted summation of the maximum likelihood reduction as in (3) and model liability quantification such as the FI would be appropriate as the new objective function. Thirdly, our proposed method relies on identifying a set of DIF-free or anchor items to identify the DIF items. In the case study, the initial targets are estimated assuming that all the items are DIF-free, and then utilized for DIF detection. However, this approach might be prone to bias when the influence of DIF on the initial trait estimation is significant. In the future, we are interested in more sophisticated DIF detection methods such as item purification with stepwise model selection (Candell and Drasgow, 1988; Kopf et al., 2015a,b). In addition, we can bypass the tedious iterative purification procedure by employing methods similar to the covariate-adjusted model with regularization (Wang et al., 2023a; Ouyang et al., 2024), where the anchor item identification and latent trait estimation are carried out simultaneously. However, model identifiability must be carefully examined, as existing methods in the literature cannot be directly applied to our setting. Last but not least, process features are

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

extracted from the action sequences and then utilized to linearly model the nuisance trait. However, a non-linear relationship between the nuisance trait and process data can be approximated by a neural network. Furthermore, it is generally difficult to systematically obtain interpretations of the nuisance traits. This is mostly because those traits are often unique to items and Z and thus not repeatedly measurable by multiple items. This is one of the limitation of the current method.

6. Acknowledgement

ChatGPT is used to polish certain parts of the writing. The authors take responsibility for the content of the paper.

7. Financial Support

This research is supported in part by NSF SES-2119938.

8. Competing interests

The authors declare none.

Appendix A: Mathematical Derivations

Proof of Proposition 1. Recall that $\mathbf{Y}^\dagger, \mathbf{Z}^\dagger, \mathbf{X}^\dagger$ are respectively the residuals of $\mathbf{Y}, \mathbf{Z}, \mathbf{X}$ after regressing on $(1, \boldsymbol{\theta})$. Regress $\mathbf{Z}^\dagger$ on $\boldsymbol{\eta} = \mathbf{X}^\dagger\boldsymbol{\omega}$:

$$\mathbf{Z}^\dagger = \beta\boldsymbol{\eta} + \boldsymbol{\epsilon}_Z,$$

and obtain the residuals $\widehat{\boldsymbol{\epsilon}}_Z := \mathbf{Z}^\dagger - \widehat{\beta}\boldsymbol{\eta}$ with $\widehat{\beta}$ as the ordinary least squares (OLS) estimate. Specifically, since $\boldsymbol{\eta}^\top\boldsymbol{\eta} = 1$ with $\mathbf{X}^{\dagger\top}\mathbf{X}^\dagger = \mathbf{I}_K$ and $\boldsymbol{\omega}^\top\boldsymbol{\omega} = 1$, we have

$$\begin{aligned} \widehat{\beta} &= \frac{\mathbf{Z}^{\dagger\top}\boldsymbol{\eta}}{\boldsymbol{\eta}^\top\boldsymbol{\eta}} = \mathbf{Z}^{\dagger\top}\boldsymbol{\eta}, \\ \widehat{\boldsymbol{\epsilon}}_Z &= \mathbf{Z}^\dagger - \boldsymbol{\eta}\mathbf{Z}^{\dagger\top}\boldsymbol{\eta} \end{aligned} \tag{A1}$$

Similarly, obtain the OLS estimates for (7):

$$\begin{aligned} \widehat{a}_1' &= \frac{\mathbf{Y}^{\dagger\top}\boldsymbol{\eta}}{\boldsymbol{\eta}^\top\boldsymbol{\eta}} = \mathbf{Y}^{\dagger\top}\boldsymbol{\eta}, \\ \widehat{\boldsymbol{\epsilon}} &= \mathbf{Y}^\dagger - \boldsymbol{\eta}\mathbf{Y}^{\dagger\top}\boldsymbol{\eta}, \end{aligned} \tag{A2}$$

and the residuals $\widehat{\boldsymbol{\delta}} = \mathbf{Y}^\dagger - \widehat{a}_1\boldsymbol{\eta} - \widehat{\lambda}\mathbf{Z}^\dagger$ for (8).

It is straightforward to show that

$$\widehat{\boldsymbol{\epsilon}} = \widehat{\lambda}\widehat{\boldsymbol{\epsilon}}_Z + \widehat{\boldsymbol{\delta}}, \quad \widehat{\lambda} = \frac{\widehat{\boldsymbol{\epsilon}}^\top\widehat{\boldsymbol{\epsilon}}_Z}{\|\widehat{\boldsymbol{\epsilon}}_Z\|^2}.$$

Therefore,

$$\|\widehat{\boldsymbol{\epsilon}}\|^2 - \|\widehat{\boldsymbol{\delta}}\|^2 = \widehat{\lambda}^2\|\widehat{\boldsymbol{\epsilon}}_Z\|^2 = \frac{(\widehat{\boldsymbol{\epsilon}}^\top\widehat{\boldsymbol{\epsilon}}_Z)^2}{\|\widehat{\boldsymbol{\epsilon}}_Z\|^2}.$$

To obtain the zero of (9) is equivalent to obtaining the zero of $\widehat{\boldsymbol{\epsilon}}^\top\widehat{\boldsymbol{\epsilon}}_Z$. Denote $\widehat{\mathbf{Y}} = \mathbf{X}^{\dagger\top}\mathbf{Y}^\dagger \in \mathbb{R}^K$ and $\widehat{\mathbf{Z}} = \mathbf{X}^{\dagger\top}\mathbf{Z}^\dagger \in \mathbb{R}^K$. With $\boldsymbol{\eta} = \mathbf{X}^\dagger\boldsymbol{\omega}, \mathbf{X}^{\dagger\top}\mathbf{X}^\dagger = \mathbf{I}_K$, $\|\boldsymbol{\eta}\| = \|\boldsymbol{\omega}\| = 1$, plugging in (A1), (A2)

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

yields

$$\begin{aligned}
\widehat{\boldsymbol{\epsilon}}^\top \widehat{\boldsymbol{\epsilon}}_Z &= (\mathbf{Y}^\dagger - \mathbf{X}^\dagger \boldsymbol{\omega} \mathbf{Y}^{\dagger\top} \mathbf{X}^\dagger \boldsymbol{\omega})^\top (\mathbf{Z}^\dagger - \mathbf{X}^\dagger \boldsymbol{\omega} \mathbf{Z}^{\dagger\top} \mathbf{X}^\dagger \boldsymbol{\omega}) \\
&= \mathbf{Y}^{\dagger\top} \mathbf{Z}^\dagger - \widehat{\mathbf{Y}}^\top \boldsymbol{\omega} \widehat{\mathbf{Z}}^\top \boldsymbol{\omega} - \boldsymbol{\omega}^\top \widehat{\mathbf{Y}}^\top \boldsymbol{\omega}^\top \widehat{\mathbf{Z}}^\top + \boldsymbol{\omega}^\top \widehat{\mathbf{Y}} \boldsymbol{\omega}^\top \boldsymbol{\omega} \widehat{\mathbf{Z}}^\top \boldsymbol{\omega} \\
&= \mathbf{Y}^{\dagger\top} \mathbf{Z}^\dagger - \boldsymbol{\omega}^\top \widehat{\mathbf{Y}} \widehat{\mathbf{Z}}^\top \boldsymbol{\omega} \\
&= \boldsymbol{\omega}^\top \left(\mathbf{Y}^{\dagger\top} \mathbf{Z}^\dagger \cdot \mathbf{I}_K - \frac{\widehat{\mathbf{Y}} \widehat{\mathbf{Z}}^\top + \widehat{\mathbf{Z}} \widehat{\mathbf{Y}}^\top}{2} \right) \boldsymbol{\omega} \\
&:= \boldsymbol{\omega}^\top \mathbf{A} \boldsymbol{\omega},
\end{aligned}$$

where $\mathbf{A} := \mathbf{Y}^{\dagger\top} \mathbf{Z}^\dagger \cdot \mathbf{I}_K - \frac{\widehat{\mathbf{Y}} \widehat{\mathbf{Z}}^\top + \widehat{\mathbf{Z}} \widehat{\mathbf{Y}}^\top}{2}$. In the above calculation, notice that $\boldsymbol{\omega}^\top \widehat{\mathbf{Y}}^\top$ and $\boldsymbol{\omega}^\top \widehat{\mathbf{Z}}^\top$ are numbers and therefore we could exchange order and take transpose when necessary. Consider the eigenvalue decomposition of $\mathbf{A}$: $\mathbf{A} = \mathbf{Q} \mathbf{S} \mathbf{Q}^\top$, where $\mathbf{Q} = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_K) \in \mathbb{R}^{K \times K}$, $\mathbf{S} = \text{diag}(s_1, \ldots, s_K)$, and $\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}_K$. By the property of matrix $\mathbf{A}$, there are closed-form expressions of its two eigen vectors and K eigenvalues. Specifically,

$$\begin{gathered}
\boldsymbol{q}_1 = \frac{\|\widehat{\mathbf{Y}}\| \widehat{\mathbf{Z}} + \|\widehat{\mathbf{Z}}\| \widehat{\mathbf{Y}}}{\left\| \|\widehat{\mathbf{Y}}\| \widehat{\mathbf{Z}} + \|\widehat{\mathbf{Z}}\| \widehat{\mathbf{Y}} \right\|}, \ \boldsymbol{q}_2 = \frac{\|\widehat{\mathbf{Y}}\| \widehat{\mathbf{Z}} - \|\widehat{\mathbf{Z}}\| \widehat{\mathbf{Y}}}{\left\| \|\widehat{\mathbf{Y}}\| \widehat{\mathbf{Z}} - \|\widehat{\mathbf{Z}}\| \widehat{\mathbf{Y}} \right\|}, \\
s_1 = \mathbf{Y}^{\dagger\top} \mathbf{Z}^\dagger - \frac{\|\widehat{\mathbf{Y}}\| \|\widehat{\mathbf{Z}}\| + \widehat{\mathbf{Y}}^\top \widehat{\mathbf{Z}}}{2}, \ s_2 = \mathbf{Y}^{\dagger\top} \mathbf{Z}^\dagger + \frac{\|\widehat{\mathbf{Y}}\| \|\widehat{\mathbf{Z}}\| - \widehat{\mathbf{Y}}^\top \widehat{\mathbf{Z}}}{2}, \\
s_3 = \ldots = s_K = \mathbf{Y}^{\dagger\top} \mathbf{Z}^\dagger.
\end{gathered}$$

Under assumption (10), we have $s_1 < 0 < s_3 = \cdots = s_K < s_2$. Let

$$\alpha = \sqrt{\frac{s_2}{s_2 - s_1}}, \quad \beta = \sqrt{\frac{-s_1}{s_2 - s_1}}$$

and

$$\widehat{\boldsymbol{\omega}} = \alpha \boldsymbol{q}_1 + \beta \boldsymbol{q}_2 \text{ or } \alpha \boldsymbol{q}_1 - \beta \boldsymbol{q}_2,$$

then we have $\|\widehat{\boldsymbol{\omega}}\| = 1$ and $L(\widehat{\boldsymbol{\omega}}) = 0$.

We notice that $L(\boldsymbol{\omega})$ has more than one zeros. We choose $\widehat{\boldsymbol{\omega}}_1 := \alpha \boldsymbol{q}_1 + \beta \boldsymbol{q}_2$ over $\widehat{\boldsymbol{\omega}}_2 := \alpha \boldsymbol{q}_1 - \beta \boldsymbol{q}_2$ because of the following geometric interpretations. We further write the scaled versions of $\widehat{\mathbf{Y}}, \widehat{\mathbf{Z}}$ as $\widetilde{\mathbf{Y}} = \widehat{\mathbf{Y}} / \|\widehat{\mathbf{Y}}\|, \widetilde{\mathbf{Z}} = \widehat{\mathbf{Z}} / \|\widehat{\mathbf{Z}}\|$. Without loss of generality, we have assumed that

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

$\mathbf{Y}^{\dagger\top}\mathbf{Z}^{\dagger} > 0$. We consider the case where $\mathbf{X}^{\dagger}$ can almost perfectly predict $\mathbf{Y}^{\dagger}$. Then $\mathbf{X}^{\dagger}\mathbf{X}^{\dagger\top}\mathbf{Y}^{\dagger} \approx \mathbf{Y}^{\dagger}$, which leads to

$$\widehat{\mathbf{Y}}^{\top}\widehat{\mathbf{Z}} = \mathbf{Y}^{\dagger\top}\mathbf{X}^{\dagger}\mathbf{X}^{\dagger\top}\mathbf{Z}^{\dagger} \approx \mathbf{Y}^{\dagger\top}\mathbf{Z}^{\dagger} > 0. \tag{A3}$$

Denote the angle between $\widehat{\mathbf{Y}}, \widehat{\mathbf{Z}}$ as Δ, which is also approximately the angle between $\mathbf{Y}^{\dagger}, \mathbf{Z}^{\dagger}$. By assumption, $\Delta \in (0, \pi/2)$.

Rewrite $\boldsymbol{q}_1, \boldsymbol{q}_2$ as

$$\boldsymbol{q}_1 = \frac{\widetilde{\mathbf{Y}} + \widetilde{\mathbf{Z}}}{\left\|\widetilde{\mathbf{Y}} + \widetilde{\mathbf{Z}}\right\|}, \quad \boldsymbol{q}_2 = \frac{\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}}{\left\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\right\|}.$$

By (A3), we also notice that

$$\frac{-s_1}{\|\widehat{\mathbf{Y}}\|\|\widehat{\mathbf{Z}}\|} \approx \frac{-\widehat{\mathbf{Y}}^{\top}\widehat{\mathbf{Z}}}{\|\widehat{\mathbf{Y}}\|\|\widehat{\mathbf{Z}}\|} + \frac{1}{2} + \frac{-\widehat{\mathbf{Y}}^{\top}\widehat{\mathbf{Z}}}{2\|\widehat{\mathbf{Y}}\|\|\widehat{\mathbf{Z}}\|} = \frac{1-\cos(\Delta)}{2} = \sin^2(\frac{\Delta}{2}),$$

$$\frac{s_2}{\|\widehat{\mathbf{Y}}\|\|\widehat{\mathbf{Z}}\|} \approx \frac{\widehat{\mathbf{Y}}^{\top}\widehat{\mathbf{Z}}}{\|\widehat{\mathbf{Y}}\|\|\widehat{\mathbf{Z}}\|} + \frac{1}{2} - \frac{-\widehat{\mathbf{Y}}^{\top}\widehat{\mathbf{Z}}}{2\|\widehat{\mathbf{Y}}\|\|\widehat{\mathbf{Z}}\|} = \frac{1+\cos(\Delta)}{2} = \cos^2(\frac{\Delta}{2}).$$

Therefore,

$$\widehat{\boldsymbol{\omega}}_1 = \alpha\boldsymbol{q}_1 + \beta\boldsymbol{q}_2 \propto \cos(\frac{\Delta}{2})\boldsymbol{q}_1 + \sin(\frac{\Delta}{2})\boldsymbol{q}_2,$$
$$\widehat{\boldsymbol{\omega}}_2 = \alpha\boldsymbol{q}_1 - \beta\boldsymbol{q}_2 \propto \cos(\frac{\Delta}{2})\boldsymbol{q}_1 - \sin(\frac{\Delta}{2})\boldsymbol{q}_2.$$

Therefore, $\widehat{\boldsymbol{\omega}}_1$ aligns with $\widetilde{\mathbf{Z}}$ and $\widehat{\boldsymbol{\omega}}_2$ aligns with $\widetilde{\mathbf{Y}}$; see Figure 7 for geometric interpretations. Furthermore, $\mathbf{X}^{\dagger}\widehat{\boldsymbol{\omega}}_1$ aligns with $\mathbf{X}^{\dagger}\mathbf{X}^{\dagger\top}\mathbf{Z}^{\dagger}$, which is the projection of $\mathbf{Z}^{\dagger}$ onto the column space of $\mathbf{X}^{\dagger}$. Similarly, $\mathbf{X}^{\dagger}\widehat{\boldsymbol{\omega}}_2$ aligns with the prediction of $\mathbf{Y}^{\dagger}$ by $\mathbf{X}^{\dagger}$. We choose to select $\widehat{\boldsymbol{\omega}}_1$ as the goal is to reduce DIF instead of using the process features to predict the response.

□

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

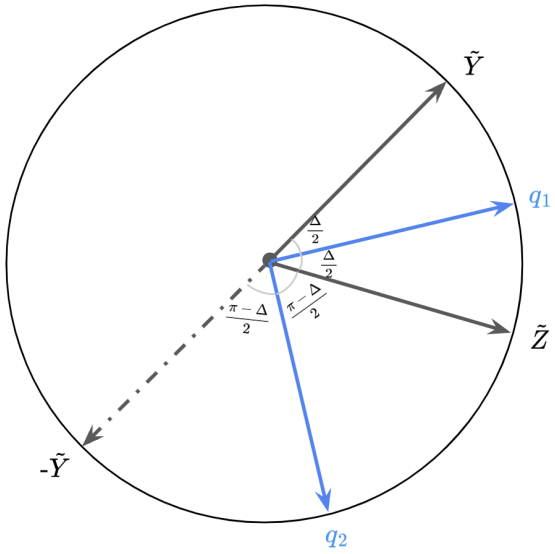

FIGURE 7.

Geometric illustration for the proof of Proposition 1.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

Appendix B: Additional Tables and Figures

B.1. Simulation with Uniform DIF

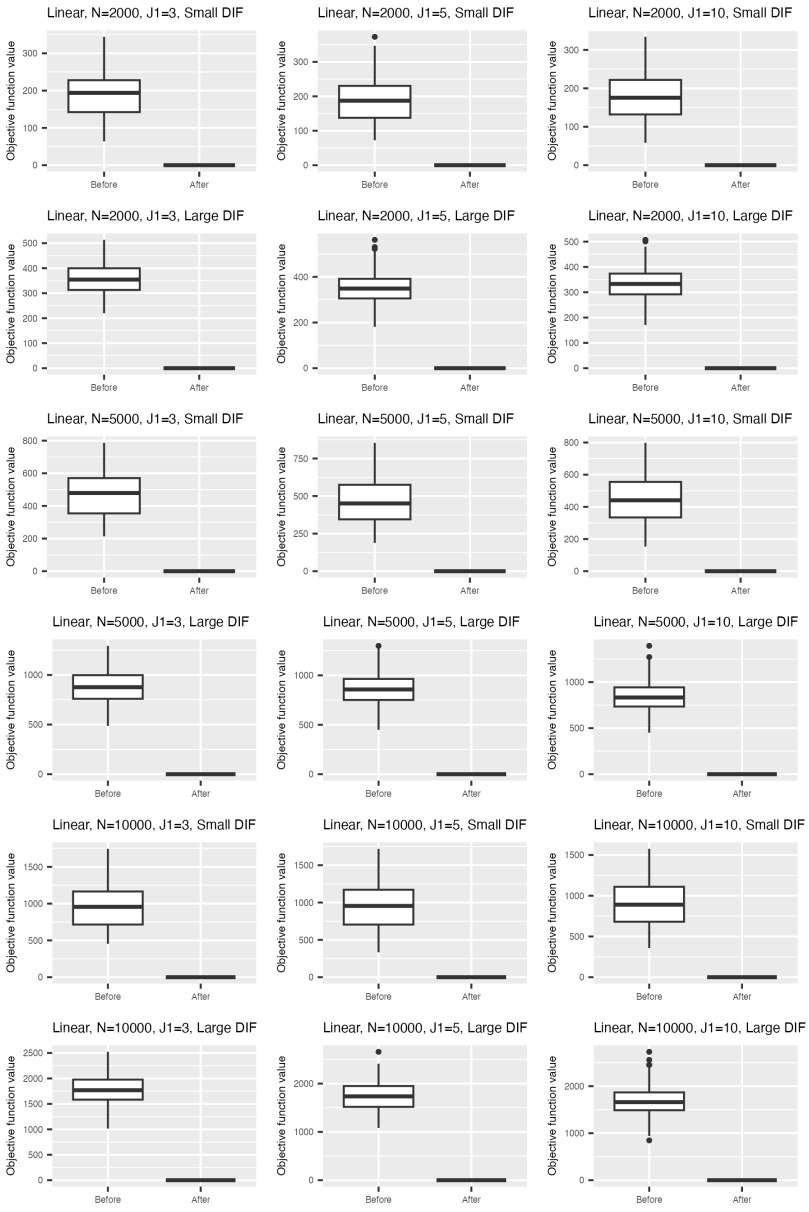

FIGURE 8.

Objective function value before and after adding the nuisance trait surrogate for the *linear model* with *uniform* DIF, under different simulation settings.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

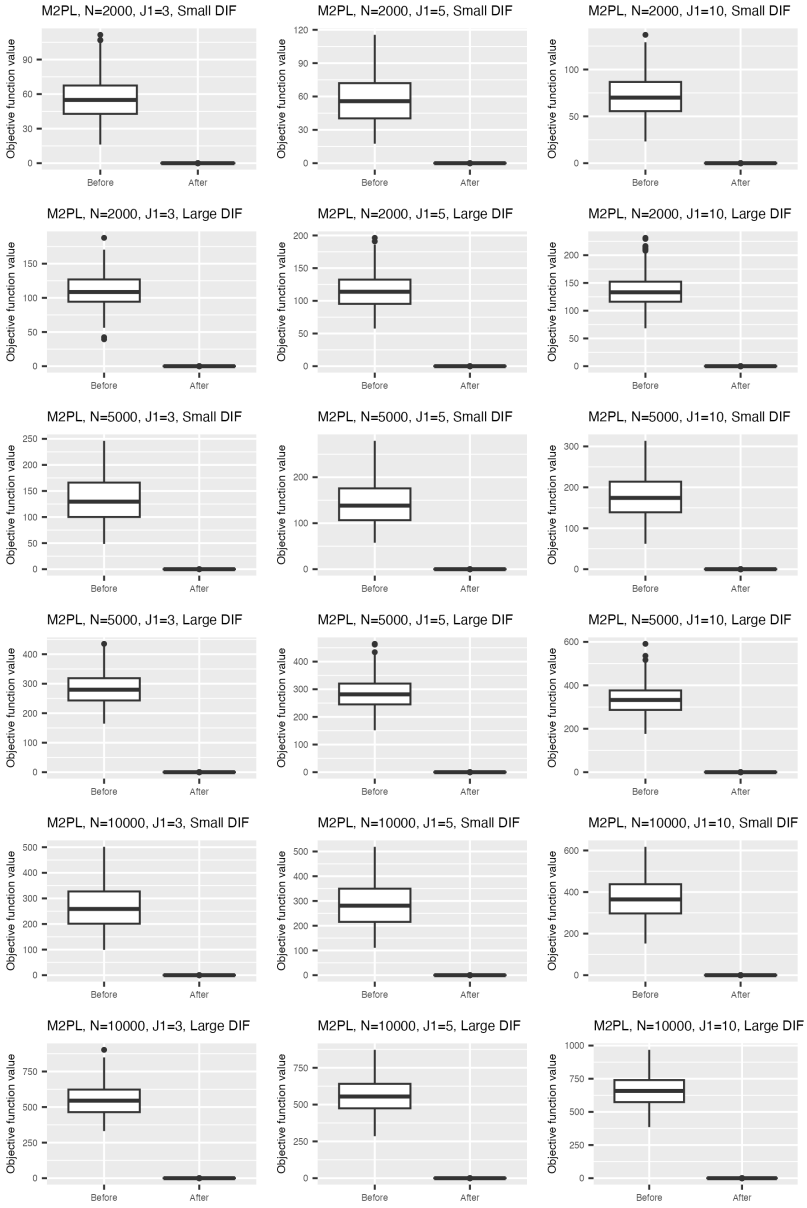

FIGURE 9.

Objective function value before and after adding the nuisance trait surrogate for the *M2PL model* with *uniform* DIF, under different simulation settings.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

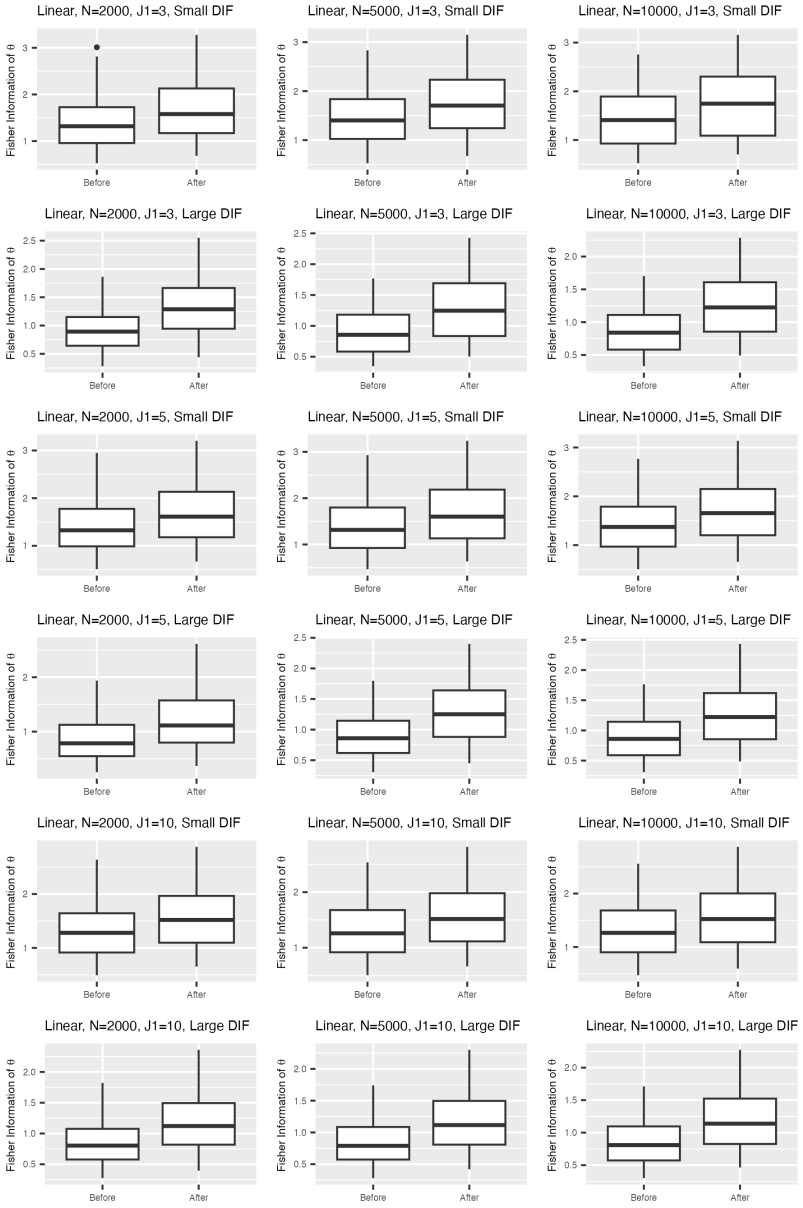

FIGURE 10.

Fisher information of θ in the measurement model before and after adding the nuisance trait surrogate for the *linear* model with *uniform* DIF, under different simulation settings.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

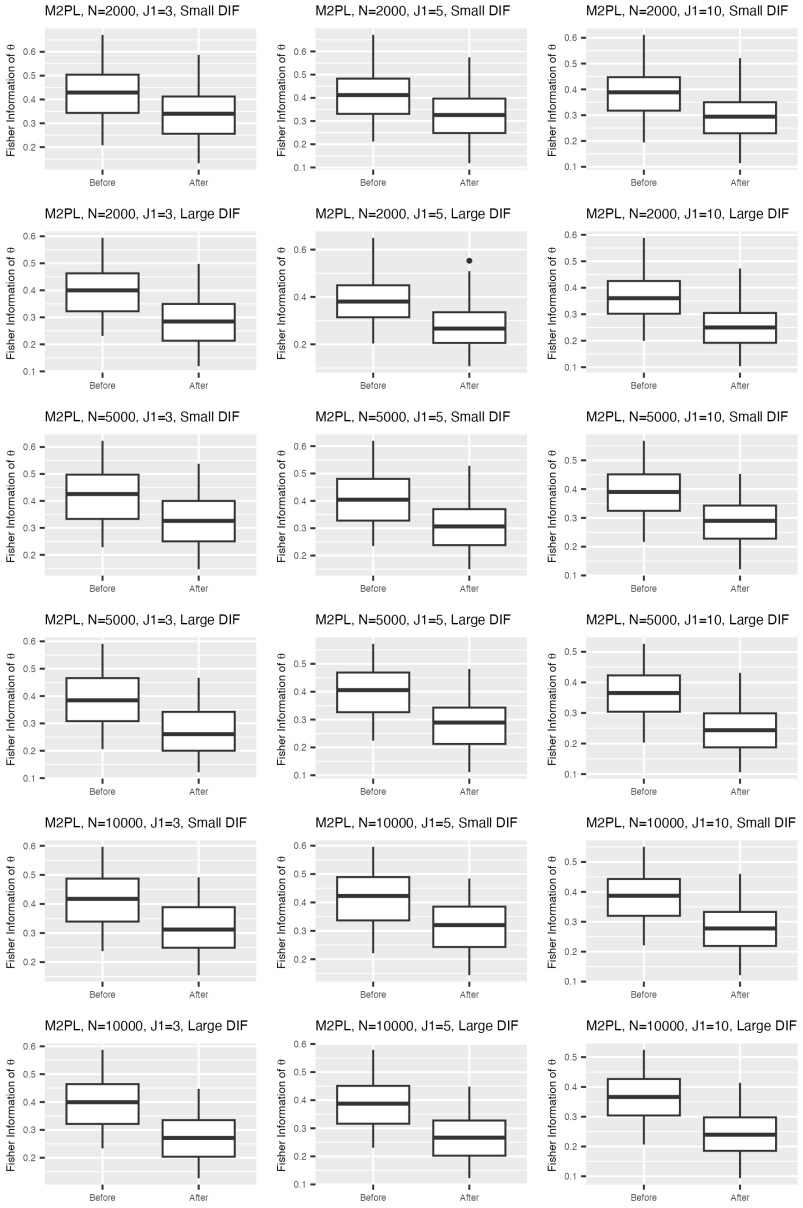

FIGURE 11.

Fisher information of θ in the measurement model before and after adding the nuisance trait surrogate for the *M2PL* model with *uniform* DIF, under different simulation settings.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

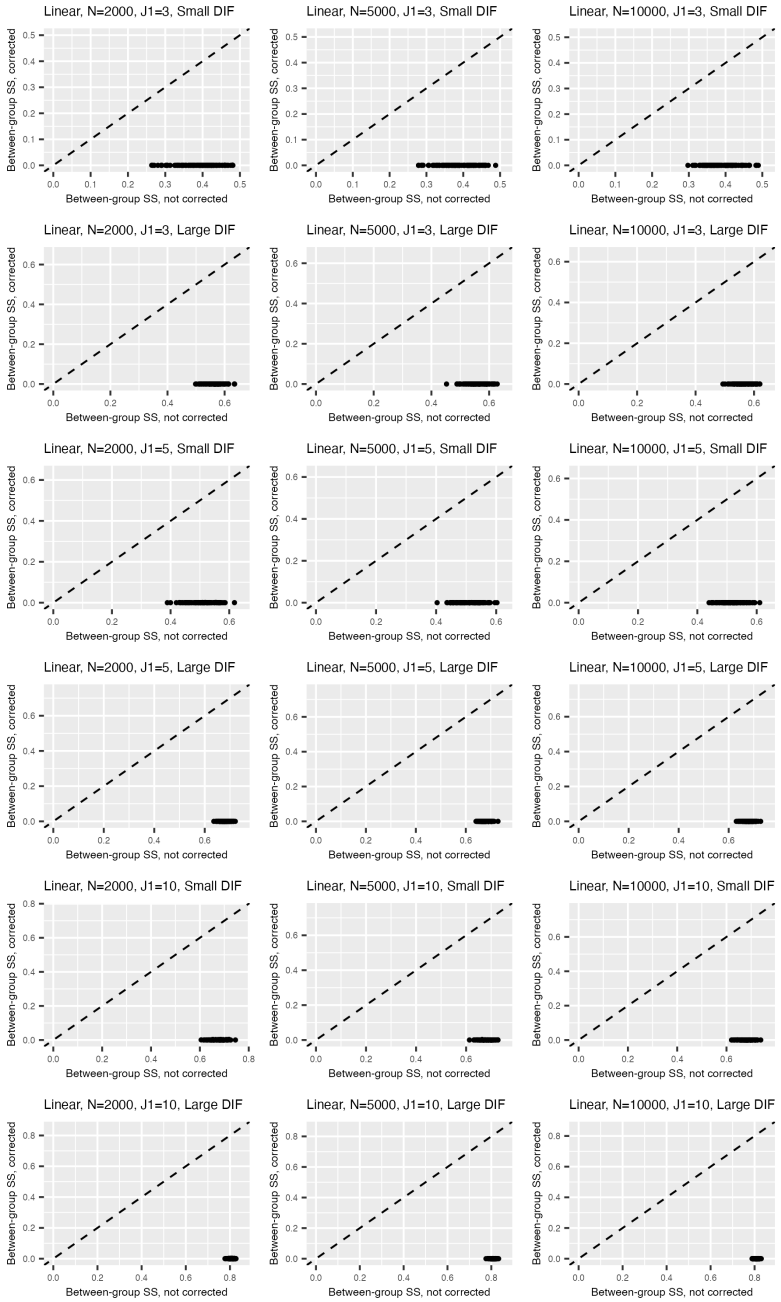

FIGURE 12.

Between-group sum of squared bias for target trait estimation for the *linear* model with *uniform* DIF, under different simulation settings. The x-axis corresponds to the estimation without DIF correction using the DIF items; the y-axis corresponds to the DIF-corrected estimation using the DIF items.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

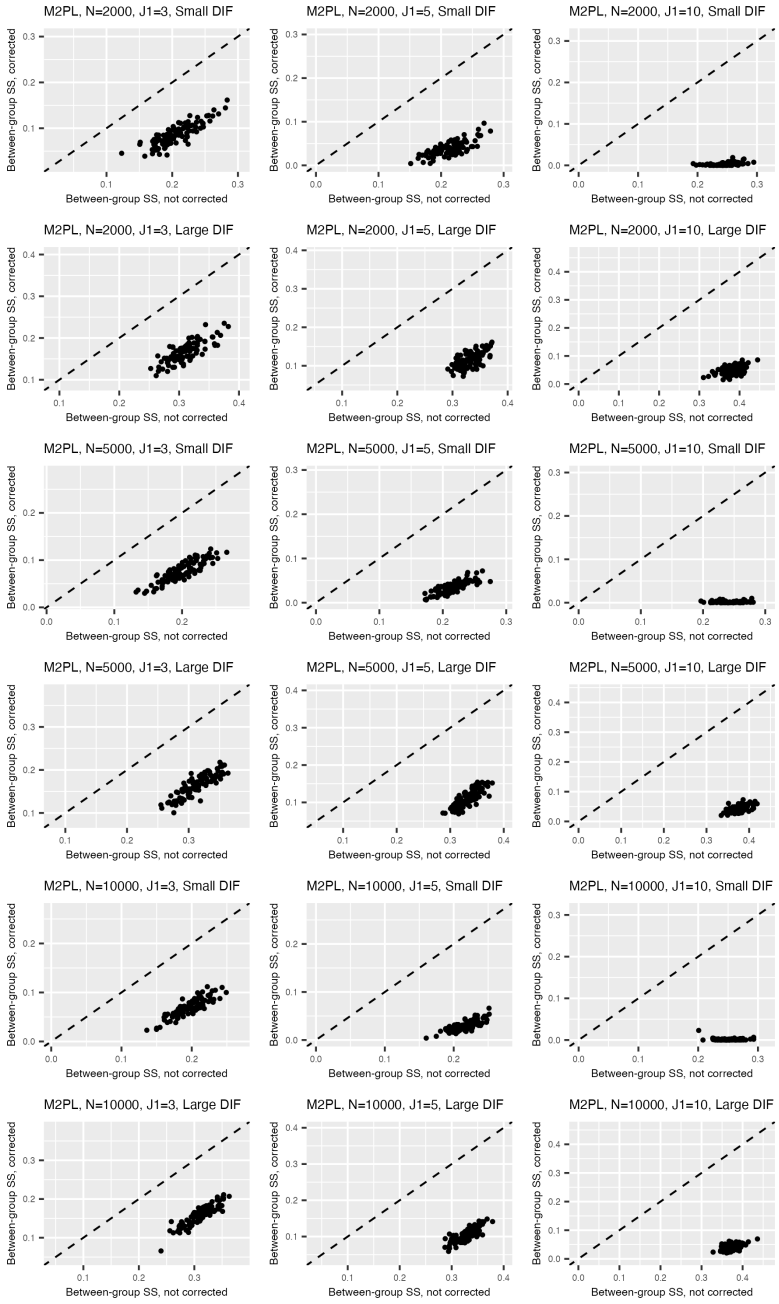

FIGURE 13.

Between-group sum of squared bias for target trait estimation for the *M2PL* model with *uniform* DIF, under different simulation settings. The x-axis corresponds to the estimation without DIF correction using the DIF items; the y-axis corresponds to the DIF-corrected estimation using the DIF items.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

B.2. Simulation with Non-uniform DIF

		Small DIF			Large DIF		
		J_1=3	J_1=5	J_1=10	J_1=3	J_1=5	J_1=10
MSE (d)	$N = 2000$	0.0325	0.0356	0.0328	0.0875	0.0886	0.0898
	$N = 5000$	0.0340	0.0319	0.0319	0.0882	0.0859	0.0882
	$N = 10000$	0.0328	0.0327	0.0319	0.0887	0.0859	0.0880
MSE (a_0)	$N = 2000$	0.0150	0.0142	0.0139	0.0380	0.0392	0.0395
	$N = 5000$	0.0147	0.0137	0.0146	0.0384	0.0398	0.0371
	$N = 10000$	0.0137	0.0135	0.0142	0.0384	0.0376	0.0362
MSE (a_1)	$N = 2000$	0.0682	0.0684	0.0658	0.1800	0.1772	0.1780
	$N = 5000$	0.0671	0.0641	0.0651	0.1783	0.1787	0.1777
	$N = 10000$	0.0639	0.0656	0.0648	0.1772	0.1762	0.1733
Corr ($\widehat{\boldsymbol{\eta}}, \boldsymbol{\eta}$)	$N = 2000$	0.6669	0.6667	0.6668	0.6628	0.6687	0.6684
	$N = 5000$	0.6684	0.6674	0.6679	0.6675	0.6676	0.6678
	$N = 10000$	0.6714	0.6694	0.6688	0.6687	0.6682	0.6710

TABLE 6.

Mean squared error of item parameter estimates and nuisance trait correlation for the *linear model* with *non-uniform* DIF under different simulation settings. The values are averaged across the DIF items and replications.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

		Small DIF			Large DIF		
		J_1=3	J_1=5	J_1=10	J_1=3	J_1=5	J_1=10
MSE (d)	$N = 2000$	0.0451	0.0471	0.0331	0.0500	0.0487	0.0381
	$N = 5000$	0.0420	0.0386	0.0287	0.0438	0.0414	0.0335
	$N = 10000$	0.0403	0.0367	0.0273	0.0411	0.0401	0.0321
MSE (a_0)	$N = 2000$	0.0153	0.0142	0.0440	0.0200	0.0187	0.0469
	$N = 5000$	0.0097	0.0122	0.0465	0.0203	0.0204	0.0575
	$N = 10000$	0.0078	0.0136	0.0478	0.0202	0.0237	0.0568
MSE (a_1)	$N = 2000$	0.0360	0.0387	0.0436	0.0561	0.0605	0.0511
	$N = 5000$	0.0147	0.0134	0.0155	0.0508	0.0497	0.0400
	$N = 10000$	0.0112	0.0104	0.0093	0.0539	0.0539	0.0395
Corr ($\widehat{\boldsymbol{\eta}}, \boldsymbol{\eta}$)	$N = 2000$	0.7400	0.7408	0.7495	0.8075	0.8023	0.8083
	$N = 5000$	0.8115	0.8161	0.8115	0.8463	0.8467	0.8465
	$N = 10000$	0.8365	0.8417	0.8366	0.8602	0.8605	0.8605

TABLE 7.

Mean squared error of item parameter estimates and nuisance trait correlation for the *M2PL model* with *non-uniform* DIF under different simulation settings. The values are averaged across the DIF items and replications.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

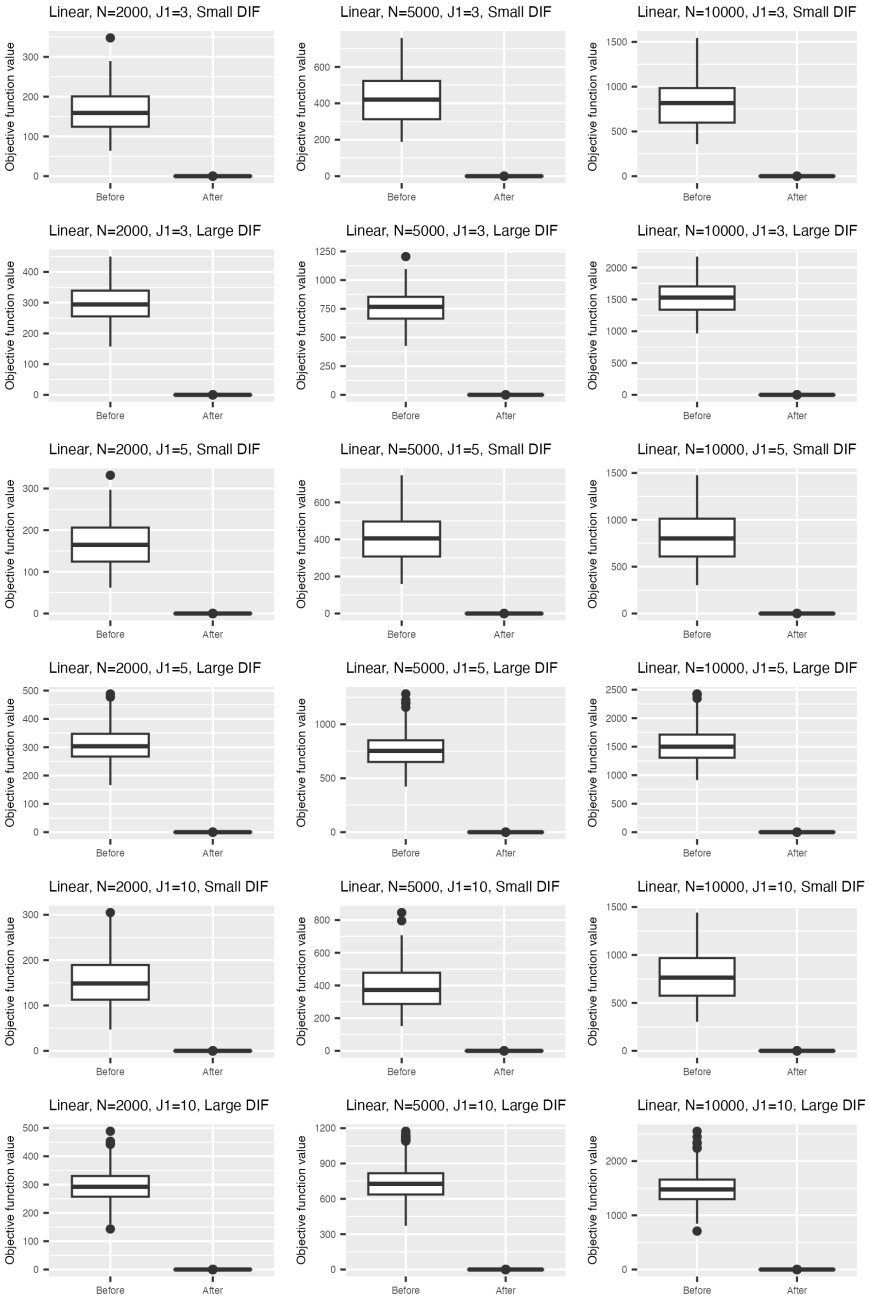

FIGURE 14.

Objective function value before and after adding the nuisance trait surrogate for the *linear model* with *non-uniform* DIF, under different simulation settings.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

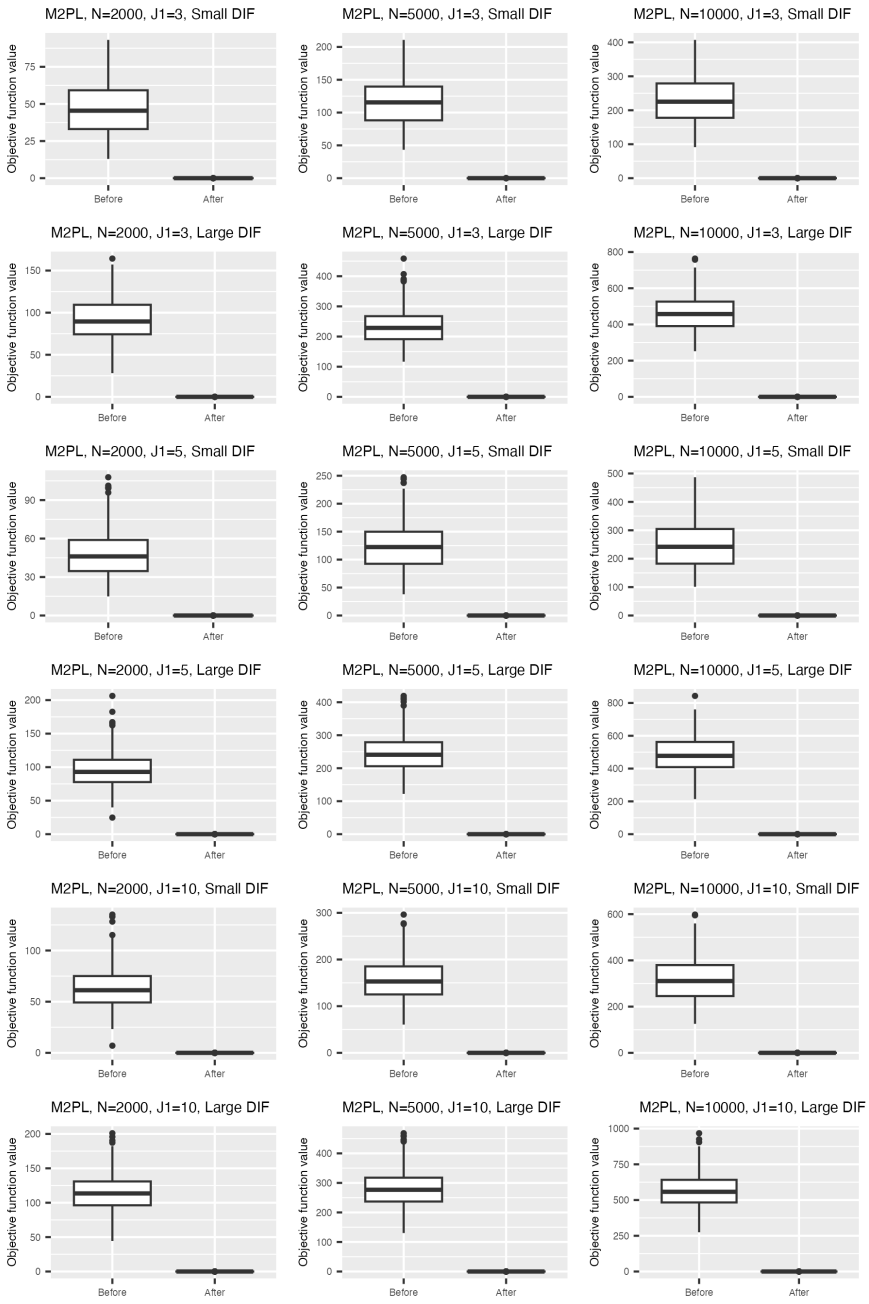

FIGURE 15.

Objective function value before and after adding the nuisance trait surrogate for the *M2PL model* with *non-uniform* DIF, under different simulation settings.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

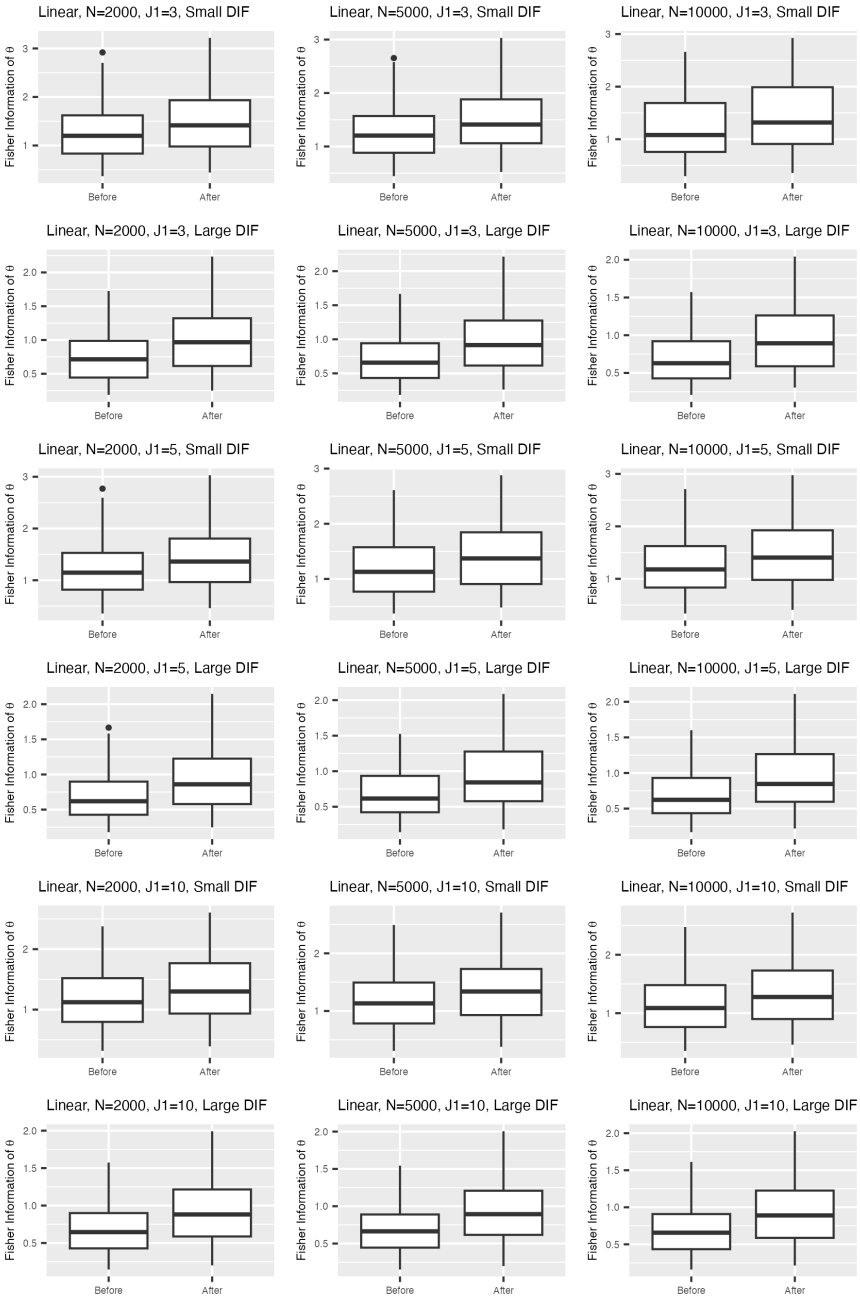

FIGURE 16.

Fisher information of θ in the measurement model before and after adding the nuisance trait surrogate for the *linear* model with *non-uniform* DIF, under different simulation settings.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

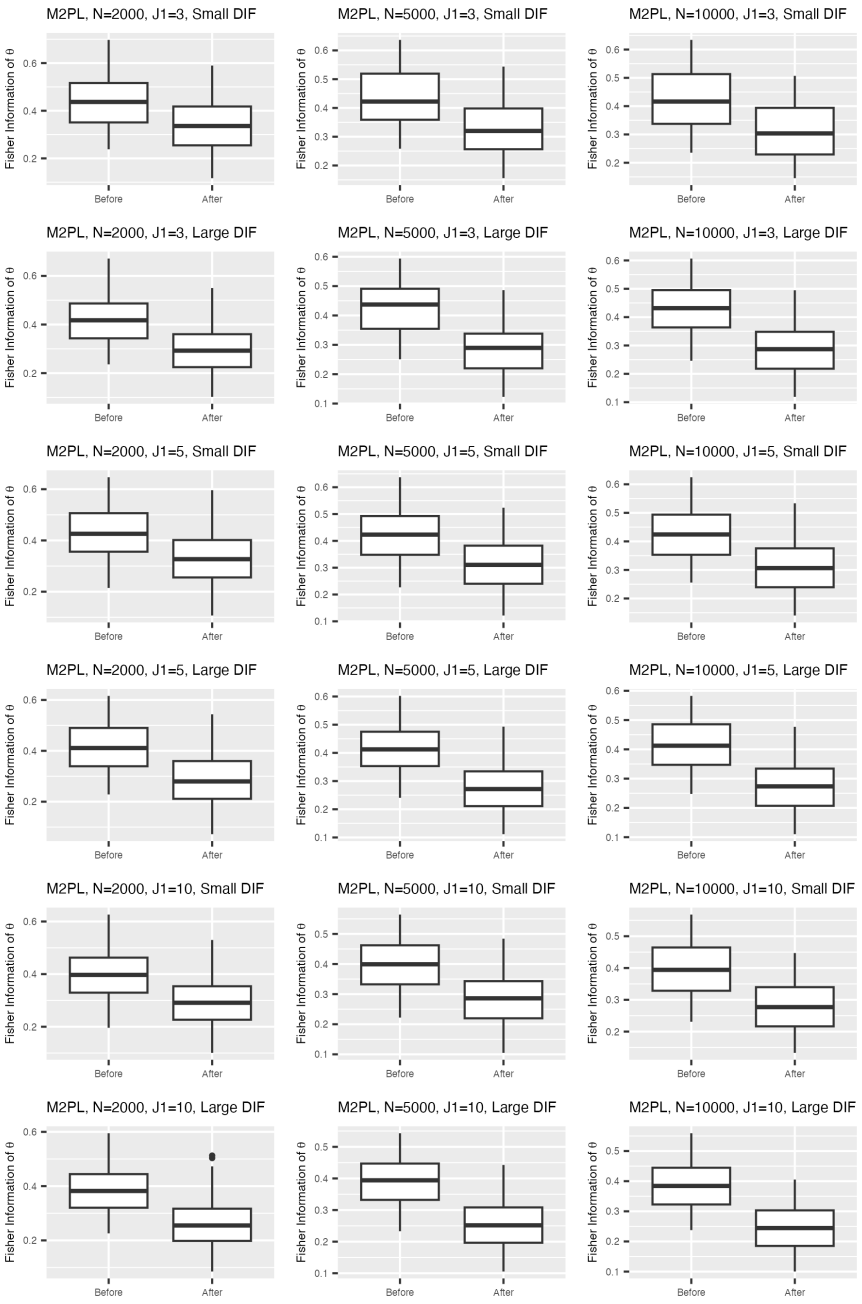

FIGURE 17.

Fisher information of θ in the measurement model before and after adding the nuisance trait surrogate for the *M2PL* model with *non-uniform* DIF, under different simulation settings.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

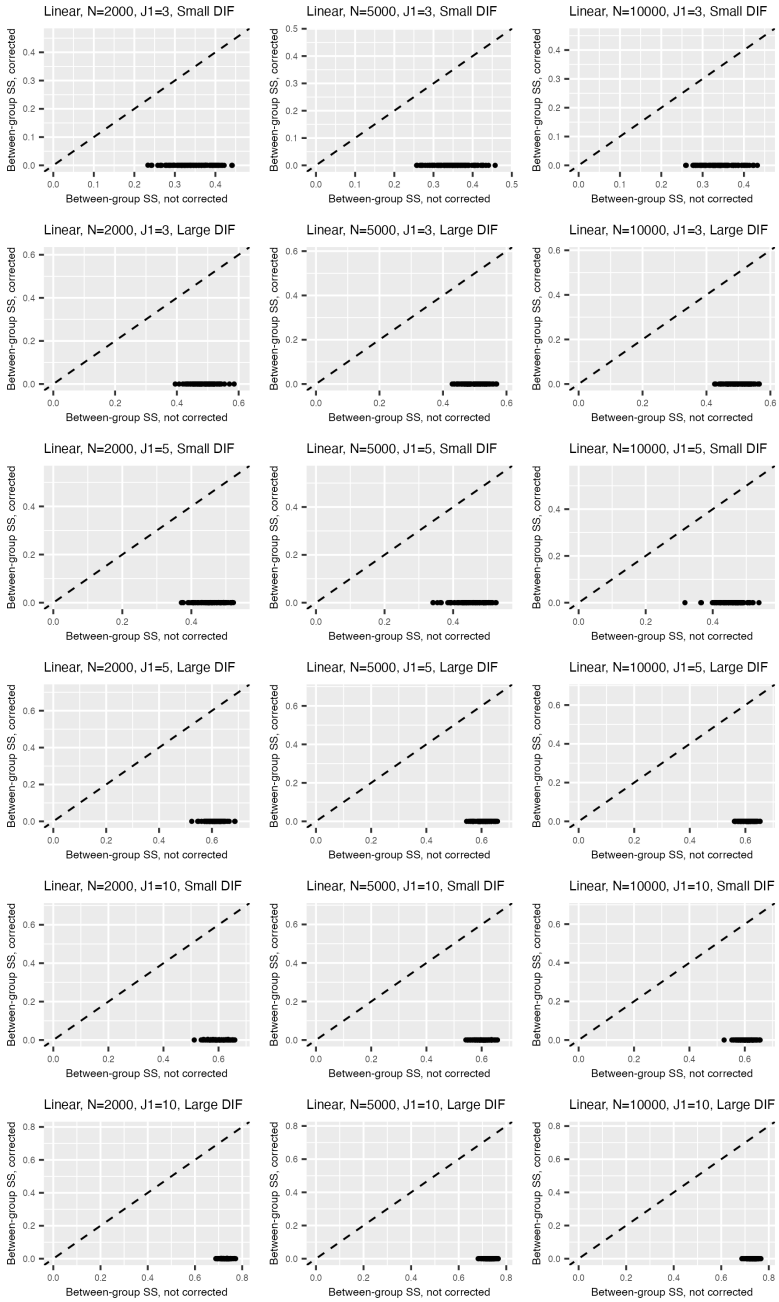

FIGURE 18.

Between-group sum of squared bias for target trait estimation for the *linear* model with *non-uniform* DIF, under different simulation settings. The x-axis corresponds to the estimation without DIF correction using the DIF items; the y-axis corresponds to the DIF-corrected estimation using the DIF items.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

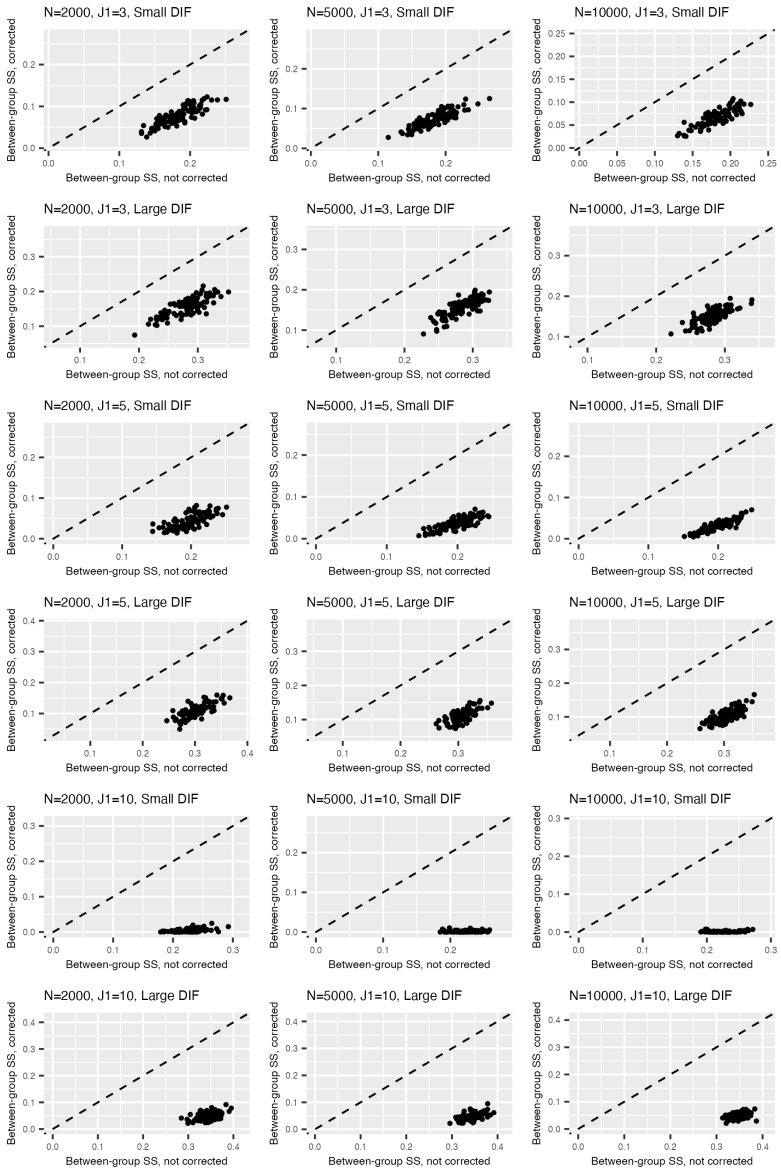

FIGURE 19.

Between-group sum of squared bias for target trait estimation for the *M2PL* model with *non-uniform* DIF, under different simulation settings. The x-axis corresponds to the estimation without DIF correction using the DIF items; the y-axis corresponds to the DIF-corrected estimation using the DIF items.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

B.3. Case Study

Item	Response mean (Response std)					
	Age		Income		Gender	
	$Z = 0$	$Z = 1$	$Z = 0$	$Z = 1$	$Z = 0$	$Z = 1$
U01a	2.33 (1.13)	1.57 (1.34)	2.18 (1.2)	2.01 (1.29)	2.15 (1.24)	2.02 (1.26)
U01b	0.60 (0.49)	0.37 (0.48)	0.56 (0.50)	0.49 (0.50)	0.55 (0.50)	0.50 (0.50)
U02	1.03 (1.18)	0.63 (1.05)	1.03 (1.19)	0.80 (1.12)	0.97 (1.18)	0.84 (1.13)
U03a	0.50 (0.50)	0.29 (0.45)	0.47 (0.50)	0.40 (0.49)	0.47 (0.50)	0.39 (0.49)
U04a	0.67 (1.18)	0.47 (1.03)	0.70 (1.19)	0.53 (1.07)	0.66 (1.17)	0.56 (1.10)
U06a	0.29 (0.45)	0.21 (0.41)	0.30 (0.46)	0.23 (0.42)	0.27 (0.44)	0.26 (0.44)
U06b	0.53 (0.50)	0.46 (0.50)	0.54 (0.50)	0.48 (0.5)	0.51 (0.50)	0.51 (0.50)
U07	0.55 (0.50)	0.45 (0.50)	0.56 (0.50)	0.48 (0.50)	0.52 (0.50)	0.52 (0.50)
U11b	1.26 (1.36)	0.72 (1.18)	1.17 (1.34)	1.02 (1.31)	1.11 (1.34)	1.06 (1.32)
U16	0.70 (0.46)	0.50 (0.50)	0.66 (0.47)	0.62 (0.49)	0.65 (0.48)	0.63 (0.48)
U19a	0.79 (0.41)	0.67 (0.47)	0.78 (0.41)	0.73 (0.45)	0.76 (0.43)	0.74 (0.44)
U19b	1.33 (0.85)	1.00 (0.91)	1.30 (0.86)	1.16 (0.9)	1.29 (0.87)	1.16 (0.89)
U23	1.63 (1.38)	1.06 (1.33)	1.52 (1.39)	1.38 (1.39)	1.50 (1.39)	1.39 (1.39)

TABLE 8.

The mean and standard deviation of the polytomous responses by group for each item. $Z = 0$ corresponds to the reference group, and $Z = 1$ corresponds to the focal group.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

Item	Response mean (Response std)					
	Age		Income		Gender	
	$Z = 0$	$Z = 1$	$Z = 0$	$Z = 1$	$Z = 0$	$Z = 1$
U01a	0.70 (0.46)	0.39 (0.49)	0.63 (0.48)	0.57 (0.49)	0.63 (0.48)	0.57 (0.50)
U01b	0.60 (0.49)	0.37 (0.48)	0.56 (0.50)	0.49 (0.50)	0.55 (0.50)	0.50 (0.50)
U02	0.18 (0.38)	0.11 (0.31)	0.19 (0.39)	0.13 (0.34)	0.17 (0.38)	0.14 (0.35)
U03a	0.50 (0.50)	0.29 (0.45)	0.47 (0.50)	0.40 (0.49)	0.47 (0.50)	0.39 (0.49)
U04a	0.17 (0.38)	0.12 (0.32)	0.18 (0.38)	0.13 (0.34)	0.17 (0.37)	0.14 (0.35)
U06a	0.29 (0.45)	0.21 (0.41)	0.30 (0.46)	0.23 (0.42)	0.27 (0.44)	0.26 (0.44)
U06b	0.53 (0.50)	0.46 (0.50)	0.54 (0.50)	0.48 (0.50)	0.51 (0.50)	0.51 (0.50)
U07	0.55 (0.50)	0.45 (0.50)	0.56 (0.50)	0.48 (0.50)	0.52 (0.50)	0.52 (0.50)
U11b	0.31 (0.46)	0.16 (0.37)	0.29 (0.45)	0.25 (0.43)	0.27 (0.45)	0.26 (0.44)
U16	0.70 (0.46)	0.50 (0.50)	0.66 (0.47)	0.62 (0.49)	0.65 (0.48)	0.63 (0.48)
U19a	0.79 (0.41)	0.67 (0.47)	0.78 (0.41)	0.73 (0.45)	0.76 (0.43)	0.74 (0.44)
U19b	0.58 (0.49)	0.42 (0.49)	0.56 (0.5)	0.50 (0.50)	0.56 (0.50)	0.49 (0.50)
U23	0.44 (0.50)	0.26 (0.44)	0.40 (0.49)	0.37 (0.48)	0.40 (0.49)	0.36 (0.48)

TABLE 9.

The mean and standard deviation of the binary responses by group for each item. $Z = 0$ corresponds to the reference group, and $Z = 1$ corresponds to the focal group. To convert the polytomous responses to binary responses, we treat the full score as correct ($Y = 1$), and other values as incorrect ($Y = 0$).

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

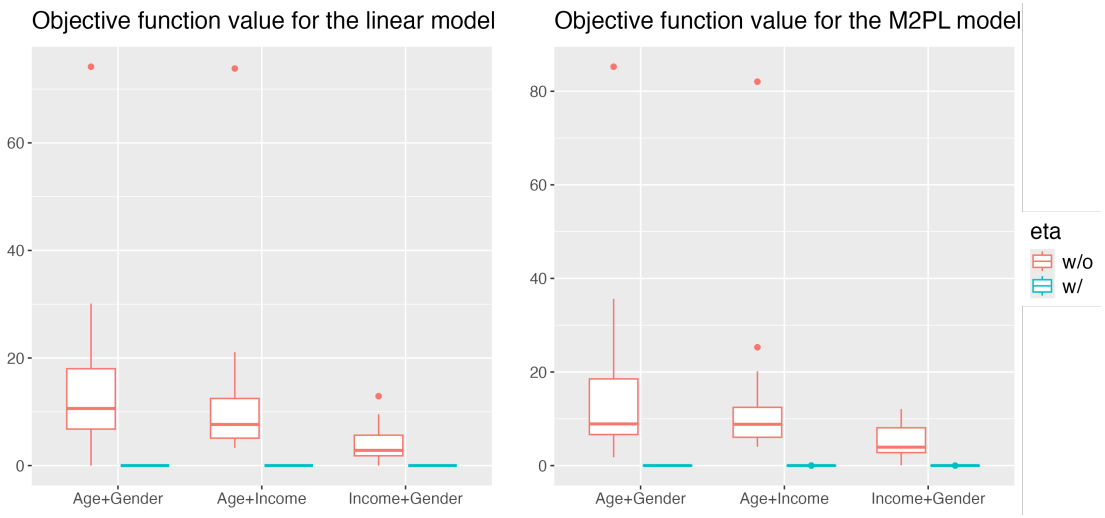

FIGURE 20.

Comparing the objective function value with and without the nuisance trait surrogate for the linear (left) and the M2PL model (right) with two grouping variables.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of educational measurement*, 29(1):67–91.

Ackerman, T. A. and Ma, Y. (2024). Examining differential item functioning from a multidimensional IRT perspective. *Psychometrika*, 89(1):4–41.

Bauer, D. J., Belzak, W. C., and Cole, V. T. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural equation modeling: a multidisciplinary journal*, 27(1):43–55.

Bergner, Y. and von Davier, A. A. (2019). Process data in naep: Past, present, and future. *Journal of Educational and Behavioral Statistics*, 44(6):706–732.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459.

Candell, G. L. and Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied psychological measurement*, 12(3):253–260.

Cao, M., Tay, L., and Liu, Y. (2017). A monte carlo study of an iterative wald test procedure for DIF analysis. *Educational and psychological measurement*, 77(1):104–118.

Chen, Y. (2020). A continuous-time dynamic choice measurement model for problem-solving process data. *psychometrika*, 85(4):1052–1075.

Chen, Y., Li, C., Ouyang, J., and Xu, G. (2023). DIF statistical inference without knowing anchoring items. *Psychometrika*, 88(4):1097–1122.

Cheng, Y., Shao, C., and Lathrop, Q. N. (2016). The mediated mimic model for understanding the underlying mechanism of DIF. *Educational and Psychological Measurement*, 76(1):43–63.

Cho, S.-J. and Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics*, 35(3):336–370.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

Cho, S.-J., Suh, Y., and Lee, W.-y. (2016a). After differential item functioning is detected: IRT item calibration and scoring in the presence of DIF. *Applied Psychological Measurement*, 40(8):573–591.

Cho, S.-J., Suh, Y., and Lee, W.-y. (2016b). An ncme instructional module on latent DIF analysis using mixture item response models. *Educational Measurement: Issues and Practice*, 35(1):48–61.

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

De Boeck, P., Cho, S.-J., and Wilson, M. (2011). Explanatory secondary dimension modeling of latent differential item functioning. *Applied Psychological Measurement*, 35(8):583–603.

Dorans, N. J. and Holland, P. W. (1992). DIF detection and description: Mantel-haenszel and standardization 1, 2. *ETS Research Report Series*, 1992(1):i–40.

Dorans, N. J. and Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of educational measurement*, 23(4):355–368.

Gao, Y., Zhai, X., Bulut, O., Cui, Y., and Sun, X. (2022). Examining humans' problem-solving styles in technology-rich environments using log file data. *Journal of Intelligence*, 10(3):38.

Halpin, P. F. (2024). Differential item functioning via robust scaling. *Psychometrika*, pages 1–26.

He, Q., Borgonovi, F., and Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Using sequence mining to identify behavioral patterns across digital tasks. *Computers & Education*, 166:104170.

He, Q., Borgonovi, F., and Suárez-Álvarez, J. (2023). Clustering sequential navigation patterns in multiple-source reading tasks with dynamic time warping method. *Journal of Computer Assisted Learning*, 39(3):719–736.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

Holland, P. W. and Thayer, D. T. (1986). Differential item functioning and the mantel-haenszel procedure. *ETS Research Report Series*, 1986(2):i–24.

Holland, P. W. and Wainer, H. (2012). *Differential item functioning*. Routledge.

Huang, Q., Bolt, D. M., and Lyu, W. (2024). Investigating item complexity as a source of cross-national dif in timss math and science. *Large-scale Assessments in Education*, 12(1):12.

Kim, S.-H., Cohen, A. S., and Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, 32(3):261–276.

Kok, F. (1988). Item bias and test multidimensionality. *Latent trait and latent class models*, pages 263–275.

Kopf, J., Zeileis, A., and Strobl, C. (2015a). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and psychological measurement*, 75(1):22–56.

Kopf, J., Zeileis, A., and Strobl, C. (2015b). A framework for anchor methods and an iterative forward approach for DIF detection. *Applied Psychological Measurement*, 39(2):83–103.

Li, Z., Shin, J., Kuang, H., and Huggins-Manley, A. C. (2024). Exploring the evidence to interpret differential item functioning via response process data. *Educational and Psychological Measurement*, page 00131644241298975.

Liang, K., Tu, D., and Cai, Y. (2023). Using process data to improve classification accuracy of cognitive diagnosis model. *Multivariate Behavioral Research*, 58(5):969–987.

Liu, X. and Jane Rogers, H. (2022). Treatments of differential item functioning: a comparison of four methods. *Educational and Psychological Measurement*, 82(2):225–253.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. *Basic problems in cross-cultural psychology*, pages 19–29.

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

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge.

OECD (2012a). *Literacy, numeracy and problem solving in technology-rich environments: Framework for the OECD survey of adult skills*. OECD publishing.

OECD (2012b). *Literacy, Numeracy and Problem Solving in Technology-Rich Environments: Framework for the OECD Survey of Adult Skills*. OECD Publishing Paris.

OECD (2012c). Results: Creative problem solving: Students' skills in tackling real-life problems, vol. v.

Ouyang, J., Chen, Y., Li, C., and Xu, G. (2025). Statistical analysis of large-scale item response data under measurement non-invariance: A statistically consistent method and its application to PISA 2022. *arXiv preprint arXiv:2505.16608*.

Ouyang, J., Cui, C., Tan, K. M., and Xu, G. (2024). Statistical inference for covariate-adjusted and interpretable generalized factor model with application to testing fairness. *arXiv preprint arXiv:2404.16745*.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53:495–502.

Roussos, L. and Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied psychological measurement*, 20(4):355–371.

Rudner, L. M., Getson, P. R., and Knight, D. L. (1980). A monte carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement*, 17(1):1–10.

Sahin, F. and Colvin, K. F. (2020). Enhancing response time thresholds with response behaviors for detecting disengaged examinees. *Large-Scale Assessments in Education*, 8:1–24.

Shan, N. and Xu, P.-F. (2024). Bayesian adaptive lasso for detecting item–trait relationship and differential item functioning in multidimensional item response theory models. *psychometrika*, 89(4):1337–1365.

Shealy, R. and Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, 58(2):159–194.

https://doi.org/10.1017/psy.2025.10072 Published online by Cambridge University Press

Swaminathan, H. and Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, 27(4):361–370.

Tang, X. (2024). A latent hidden markov model for process data. *psychometrika*, 89(1):205–240.

Tang, X., Wang, Z., He, Q., Liu, J., and Ying, Z. (2020a). Latent feature extraction for process data via multidimensional scaling. *Psychometrika*, 85(2):378–397.

Tang, X., Wang, Z., Liu, J., and Ying, Z. (2020b). An exploratory analysis of the latent structure of process data via action sequence autoencoders. *British Journal of Mathematical and Statistical Psychology*, 74.

Thissen, D., Steinberg, L., and Wainer, H. (2013). Use of item response theory in the study of group differences in trace lines. In *Test validity*, pages 147–169. Routledge.

Ulitzsch, E., He, Q., and Pohl, S. (2022). Using sequence mining techniques for understanding incorrect behavioral patterns on interactive tasks. *Journal of Educational and Behavioral Statistics*, 47(1):3–35.

Wallin, G., Chen, Y., and Moustaki, I. (2024). DIF analysis with unknown groups and anchor items. *Psychometrika*, pages 1–29.

Wang, C., Zhu, R., and Xu, G. (2023a). Using lasso and adaptive lasso to identify DIF in multidimensional 2PL models. *Multivariate behavioral research*, 58(2):387–407.

Wang, Z., Tang, X., Liu, J., and Ying, Z. (2023b). Subtask analysis of process data through a predictive model. *British Journal of Mathematical and Statistical Psychology*, 76(1):211–235.

Wise, S. L. and DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, 43(1):19–38.

Wise, S. L. and Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18(2):163–183.

Woods, C. M., Cai, L., and Wang, M. (2013). The langer-improved wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, 73(3):532–547.

Xiao, Y., He, Q., Veldkamp, B., and Liu, H. (2021). Exploring latent states of problem-solving competence using hidden markov model on process data. *Journal of Computer Assisted Learning*, 37(5):1232–1247.

Xiao, Y. and Liu, H. (2024). A state response measurement model for problem-solving process data. *Behavior Research Methods*, 56(1):258–277.

Zhang, S., Wang, Z., Qi, J., Liu, J., and Ying, Z. (2023). Accurate assessment via process data. *psychometrika*, 88(1):76–97.

Zwick, R., Thayer, D. T., and Lewis, C. (2000). Using loss functions for DIF detection: An empirical bayes approach. *Journal of Educational and Behavioral Statistics*, 25(2):225–247.