Applied Psychological Measurement

Detecting Differential Item Functioning for Continuous Response CAT

Journal:	*Applied Psychological Measurement*
Manuscript ID	APM-22-12-168
Manuscript Type:	Manuscripts
Keywords:	Continuous response model, Computerized adaptive testing, Differential item functioning

Detecting Differential Item Functioning for Continuous Response CAT

Abstract

Evaluating items for potential differential item functioning (DIF) is an essential step to ensuring measurement fairness. In this paper, we focus on a specific scenario, namely, the continuous response, severely sparse, computerized adaptive testing (CAT). Continuous responses items are growingly used in performance-based tasks because they tend to generate more information than traditional dichotomous items. Severe sparsity arises when large number of items are automatically generated via machine learning algorithms. We propose two DIF detection methods in this scenario. The first is a modified version of the CAT SIBTEST, a non-parametric method that does not depend on any specific item response theory model assumptions. The second is a regularization method, a parametric, model-based approach. Simulation studies show that both methods are effective in correctly identifying items with uniform DIF. A real data analysis is provided in the end to illustrate the utility and potential caveats of the two methods.

Key words: computerized adaptive test, DIF, continuous response, SIBTEST

Introduction

Measurement developers often examine the fairness of their instruments by studying their measure's item-factor relationships across different groups of respondents. Presumably, an individual's response on an item should depend only on the latent variable(s) the item intends to measure (e.g., understanding scientific principles); it should not depend on other construct-irrelevant person-level characteristics. However, an item is considered to exhibit differential item functioning (DIF) when two an item's characteristics differ for different groups of individuals after controlling for overall differences in performance (Holland & Thayer, 1988). Two types of DIF are often differentiated: uniform DIF and non-uniform DIF. The former refers to an item having a constant advantage for a particular group, whereas the latter refers to the advantage varying in magnitude and/or direction across the latent trait continuum (Penfield & Camilli, 2006; Woods & Grimm, 2011). When DIF is found, the validity of the measurement is called into question: the offending item will require further inspection and may need to be removed or revised.

The current operational DIF analysis is predominately based on Mantel-Haenszel (1959) chi-square statistics (MH; Holland & Thayer, 1988), simultaneous item bias test (SIBTEST, Shealy & Stout, 1993, Chang, Mazzeo, & Roussos, 1996), and an effect size based on the standardized mean difference in item scores across groups (Dorans & Kulick, 1986; Zwick, Donoghue, & Grima, 1993). All these methods are non-parametric as they do not depend on any specific item response theory (IRT) models. However, they are only designed for dichotomously or polytomously scored items, and they are more powerful in detecting uniform DIF than non-uniform DIF (Lei, Chen, & Yu, 2006). In CAT where total score is no longer a valid matching criterion, both MH and SIBTEST are modified using latent trait as the matching variable

(Nandakumar & Roussos, 1997; Zwick, Thayer, & Wingersky, 1994).

Other popular and parametric DIF detection methods are logistic regression (LR) (Swaminathan & Rogers, 1990) and IRT-likelihood ratio test (IRT-LRT; Thissen, Steinberg, & Wainer, 1993). LR models item responses as a function of group indicators, trait estimates (denoted by θ), and their interaction. Then it compares a null model (assuming no DIF) on an item to two nested models formed in hierarchy with an explanatory group variable and group-by-θ interaction variable. It does not model latent trait differences (i.e., impact) directly across groups. IRT-LRT, on the other hand, models differences in item parameters between groups conditioning on the invariance of other items in the test (i.e., anchor items), and it can model impact. Both LR and IRT-LRT are adapted in CAT (Lei, et al., 2006) and they show good Type I error control and adequate power.

Recently, a new set of methods for DIF detection starts to emerge in literature, namely, the regularization methods. Regularization is a machine learning technique that imposes a penalty function during estimation to remove parameters that have little influence on the fit of the model (Bauer, Belzak, & Cole, 2020; Belzak & Bauer, 2020; Magis, Tuerlincks, & de Boeck, 2015; Tutz & Schauberger, 2015; Wang, et al., 2021). In the context of DIF detection, an item level DIF parameter is introduced for each covariate and item parameter type. Then a penalty is imposed on the DIF parameters, and with appropriate regularization algorithms, they will either shrink to 0 implying no DIF or remain non-zero implying DIF. The advantage of the regularization method is, compared to other model-based methods reviewed above, that it does not require pre-specifying anchor items. In addition, all these studies have found that the regularization method maintains good Type I error control and reasonably high power even when there is a large proportion of DIF items, whereas LRT usually fails with severely inflated Type I error. However, the regularization methods have not been evaluated in CAT, especially severely sparse CAT.

In addition, although various methods for DIF detection have been broadly discussed and evaluated, these methods cannot readily handle continuous responses. With recent advances in educational measurement that make use of performance-based items, adaptive testing with non-categorical response items (e.g., when the examinee's performance is rated on a continuous scale between 0 and 1) is becoming more prevalent. Thus, in this paper, we propose DIF detection methods designed specifically for assessments that are growingly used in online learning, in which (1) the items are automatically generated and hence the number of total items is exceedingly large; (2) the items are administered adaptively, which means that the examinee will receive a very small subset of all potential items, and as such, there will be a high proportion of missing data/empty person-item cells; and (3) the type of response will be on a continuous scale, rather than the typical binary or Polytomous options. In what follows, we will introduce the model for continuous responses, followed by the two proposed DIF detection methods. Then we will present the simulation study and a real data analysis.

Methods

Model Description

An example of a continuous item example is a sentence completion item in an English language assessment[1]. In this task, the first and last sentences of a text are fully intact, whereas alternating words in the middle sentences are damaged by masking the second half of the word. Test takers need to rely on context and discourse information to reconstruct the damaged words. In this case, while the completion of each word can be scored as right/wrong (i.e., binary), the

[1] It is called C-test in Duolingo English Test (Cardwell, LaFlair, & Settles, 2022). Such tasks are found to be highly correlated with other major language proficiency tests and are related to spelling skills (Khodadady, 2014).

score for a paragraph is often computed as the proportion of damaged words that are correctly reconstructed. Such a score is a continuous variable bounded between 0 and 1. In Item Response Theory (IRT), the earliest continuous response model was proposed by Samejima (1973) as a limiting case of the graded response model. Wang and Zeng (1998) proposed a different parameterization of it. The primary idea is to translate a bounded response y as $z = \log\frac{y}{k-y}$ where k is the highest score (in the example above, $k = 1$) such that z is unbounded and assumed to follow a normal distribution. In this paper, we use a signed response theory model (SRT; Maris, 2020; Maris & van der Maas, 2012) instead for two reasons: (1) the model was operationally used in the real data example that will be described later, and (2) it is a continuous extension of the Rasch model and hence enjoys a nice property of the Rasch model, i.e., the score of an individual is considered to be a sufficient statistic of his/her ability. In other words, the response vector $\boldsymbol{Y}$ is independent of ability θ given the scoring rule $S(\boldsymbol{Y})$.

Based on the assumption of sufficient statistics, Maris and van der Maas (2012) derived the SRT that belongs to the exponential distribution. For person i and item j, the density function of a continuous response y_{ij} is as follows:

$$P\left(y_{ij}\middle|\theta_i\right) = \frac{\exp(y_{ij}(\theta_i - b_j))}{\int_0^1 \exp(y(\theta_i - b_j))\,dy} = \frac{\exp(y_{ij}(\theta_i - b_j))}{\dfrac{\exp(\theta_i - b_j)\text{-}1}{\theta_i - b_j}}\ , \qquad (1)$$

where b_j is the difficulty level of item j. Given the known item parameters, the ability θ_i can be estimated by simply maximizing the likelihood constructed from SRT, $l_i(\theta) = \sum_j \log(P(y_{ij}|\theta_i))$, resulting in a maximum likelihood estimate (MLE). Maris (2020) suggested to estimate θ_i by minimizing the cross entropy (MCE) defined as

$CE(\theta_i) = \sum_j\left[y_{ij}\log\left(P_i(\theta)\right) + \left(1 - y_{ij}\right)\log(1 - P_i(\theta))\right]$, where $P_i(\theta) = \frac{1}{1+\exp\text{-}(\theta_i - b_j)}$ is simply the item response function of a Rasch model. Different from the SRT likelihood, minimizing cross entropy inherently ignores the randomness of the item responses. Our preliminary study reveals that when the item responses are simulated from SRT and when test length is long (≥ 10), MLE is more accurate than MCE. However, for shorter test length (≤ 6), MCE is more accurate.

Two DIF Detection Methods

Modified CATSIB Method. This method is an extension of the CATSIB method proposed in Nandakumar and Roussos (2004) to the signed response theory model. As a variant of the SIBTEST, we will first briefly describe the SIBTEST by Shealy and Stout (1993). Let $P_R(\theta)$ and $P_F(\theta)$ denote the probabilistic item response on a given item conditional on θ, where the subscript R and F refers to reference and focal group respectively. Item subscript is omitted to avoid clutter. Then DIF on this item at a fixed value of θ is defined as

$$DIF(\theta) = P_R(\theta) - P_F(\theta). \tag{2}$$

Then DIF for an item, denoted as d, is defined as the average of $DIF(\theta)$ over the distribution of θ, i.e.,

$$d = \int DIF(\theta) f(\theta)\, d\theta, \tag{3}$$

where $f(\theta)$ denotes a density of θ based on combined focal and reference groups. Then the null and alternative hypothesis for DIF is $H_0: d = 0$ vs. $H_a: d \neq 0$.

For Equation (2) to work, the reference and focal group test takers need to be matched on ability before comparing their performance on the target item. In traditional non-adaptive tests, test takers are matched on their *estimated true score* from their *observed total score*, via a so-called regression correction approach (Shealy & Stout, 1993). In CAT, the latent trait estimate $\hat{\theta}$ serves as a natural matching criterion. To correct for potential impact (i.e., difference in θ between two groups), Nandakumar and Roussos (2004) also proposed a regression correction approach, which is theoretically equivalent to the regression correction employed in the original

SIBTEST. Their method is now modified to be suitable for the continuous scoring rubric used in SRT.

First, the test takers will be matched on $\hat{\theta}^* = E_g\left[\theta \middle| \hat{\theta}\ \right]$, where $\hat{\theta}$ is the estimated latent trait via MLE, and g is a group indicator (i.e., g= R or F). According to the derivations in Nandakumar and Roussos (2004), we have

$$E_g\left[\theta \middle| \hat{\theta}\ \right] = E_g[\theta] + \rho_g^2\left(\hat{\theta} - E_g\left[\hat{\theta}\right]\right) \tag{4}$$

where both $E_g[\theta]$ and $E_g\left[\hat{\theta}\right]$ are the mean of $\hat{\theta}$ in group g. ρ_g^2 is a reliability measure that could be obtained as follows: $\rho_g^2 = 1 - \frac{\sigma_e^2}{\sigma_{\hat{\theta}}^2}$. Here $\sigma_{\hat{\theta}}^2$ is the sample variance of $\hat{\theta}$ in group g, and σ_e^2 is the error variance from the CAT administration. Specifically, in a variable-length CAT, the test often terminates when the standard error of measurement is below a prespecified cutoff (e.g., Wang, et al., 2013, 2018). Then σ_e^2 can be obtained as the average squared standard error of $\hat{\theta}$ at the conclusion of the test.

After obtaining $\hat{\theta}^*$, the estimated DIF from Equation (3) can be approximated as

$$\hat{d} = \sum\nolimits_{q=1}^{Q}\left[\bar{P}_R\left(\hat{\theta}_q^*\right) - \bar{P}_F\left(\hat{\theta}_q^*\right)\right] p\left(\hat{\theta}_q^*\right), \tag{5}$$

where $\hat{\theta}_q^*$ is the q^{th} quadrature point selected from an interval of $[\hat{\theta}_{min}^*, \hat{\theta}_{max}^*]$. Because $\hat{\theta}^*$ is on a real-valued continuous scale, we can divide the observed $\hat{\theta}^*$ range into Q equal intervals, and $\hat{\theta}_q^*$ is the mean of the q^{th} interval. $p\left(\hat{\theta}_q^*\right)$ is the observed proportion of reference and focal group test takers in the q^{th} interval. $\bar{P}_R\left(\hat{\theta}_q^*\right)$ is the mean score of people with ability estimate in the q^{th} interval in the reference group, and $\bar{P}_F\left(\hat{\theta}_q^*\right)$ is defined similarly. The standard error for $\hat{d}$ can be estimated from the observed variance of the item responses in each ability interval, i.e.,

$$SE(\hat{d}) = \sqrt{\sum_{q=1}^{Q} \left[\frac{\hat{\sigma}_{R,q}^2(Y)}{n_{R,q}} + \frac{\hat{\sigma}_{F,q}^2(Y)}{n_{F,q}}\right] \left[p(\hat{\theta}_q^*)\right]^2}. \tag{6}$$

In Equation 6, $\hat{\sigma}_{R,q}^2(Y)$ is the observed variance of item responses from test takers in the q^{th} interval in the reference group and $\hat{\sigma}_{F,q}^2(Y)$ is defined similarly. $n_{R,q}$ and $n_{F,q}$ denote the number of test takers in the q^{th} interval in the reference and focal groups, respectively. The null hypothesis of no DIF is rejected if the following test statistic,

$$B = \frac{\hat{\beta}}{SE(\hat{\beta})}, \tag{7}$$

exceeds the $100\left(1-\frac{\alpha}{2}\right)^{th}$ percentile or below the $100\left(\frac{\alpha}{2}\right)^{th}$ percentile from the standard normal distribution.

As indicated in Nandakumar and Roussos (2004), the number of intervals may affect the performance of the test statistic. First, to ensure stable estimate of mean and variance per interval per group, there must be enough test takers in each interval, say a minimum of 5. If the count is smaller than 5 for certain interval, the cell will be eliminated. On the other hand, if the interval is too coarse, the test statistic will be overly sensitive to impact. Hence a good balance is needed. Despite of large number of Duolingo test takers, due to the large number of items and potentially small number of test takers per item, we will start with a reasonably large number of intervals, say, 60, and then monitor how many cells may be eliminated due to getting rid of sparse cells. The number of intervals will decrease gradually until no more than 5% of either the reference or focal group test takers are eliminated.

SRT Regularization Method. While the modified CATSIB method does not rely directly on estimated items parameters, the regularization method is a model-based parametric method. DIF parameters for an item (i.e., $\boldsymbol{\beta}_j$) are introduced in the density function of the continuous item response y_{ij} by modifying Equation 1 as follows:

$$L\left(y_{ij}\middle|\theta_i,\boldsymbol{\beta}_j\right)=\frac{\exp\left(y_{ij}(\theta_i-b_j+\boldsymbol{X}_i\boldsymbol{\beta}_j)\right)}{\frac{\exp\left(\theta_i-b_j+\boldsymbol{X}_i\boldsymbol{\beta}_j\right)-1}{\theta_i-b_j+\boldsymbol{X}_i\boldsymbol{\beta}_j}} \qquad (8)$$

where $\boldsymbol{X}_i$ is a 1-by-P vector including all the grouping information related to DIF for person i. For instance, if we are interested in both gender and first language that may cause DIF, then $\boldsymbol{X}_i$ will contain dummy coded information about person i's gender and their first language. $\boldsymbol{\beta}_j$ is a P-by-1 vector of regression coefficients implying the effect of grouping variables on item responses. By way of this parameterization, if item j does not have DIF, then $\boldsymbol{\beta}_j = 0$.

In contrast to CATSIB in which only two groups are compared at a time, this parameterization allows for simultaneous testing of multiple covariates as well as covariates with more than two levels. For instance, assume first language takes 3 levels: Chinese, Japanese and Korean. Then $P = 2$ in this case such that test takers with Chinese as first language will have $\boldsymbol{X}_i = (0,0)$, those with Japanese as first language will have $\boldsymbol{X}_i = (1,0)$, whereas those with Korean as first language will have $\boldsymbol{X}_i = (0,1)$.

Note that like the multiple-group IRT approach, θ_i in Equation 8 can be written as $\theta_{i(g)}$ to reflect that the distribution of θ is allowed to differ across different groups, hence it can naturally model impact, if needed. When there are DIF-free anchor items, then depending upon the combination of groups that test-takers could be exclusively assigned to, the mean and variance of $\theta_{i(g)}$ for non-reference group could be freely estimated. In this study, we intend to evaluate the comparative performance of this method versus the modified CATSIB method, only pairs of focal/reference groups will be evaluated separately. Therefore both $\boldsymbol{X}_i$ and $\boldsymbol{\beta}_j$ become scalars.

During the model estimation, a constrained least absolute shrinkage and selection operator (Lasso; Friedman, Hastie, & Tibshirani, 2010; Tibshirani, 1996) will be added on the

DIF parameter for every item. To perform regularization, we consider two methods: (1) treat everyone's θ as known from CAT and use conditional likelihood (denoted as "Lasso_Conditional" in the results section); and (2) treat θ as unknown and use marginal log-likelihood (denoted as "Lasso_Marginal").

In the conditional likelihood method, each β_j is estimated by maximizing the following objective function,

$$l(\beta_j) = \sum_{i=1}^{N_j} \log L(\boldsymbol{Y}_i|\theta_i, \beta_j) - \lambda|\beta_j| \tag{9}$$

where $L(\boldsymbol{Y}_i|\theta_i, \beta_j) = \prod_j L(y_{ij}|\theta_i, \beta_j)$ from Equation 8. This objective function can be maximized using the L-BGFS-B method available in the "optimize" function in the "stats" library of R.

In the marginal likelihood method, the log-marginal likelihood is

$$(\boldsymbol{\beta}, \boldsymbol{\Delta}) = \sum_{i=1}^{N_j} log \int_{\boldsymbol{\theta}_i} L(\boldsymbol{Y}_i|\theta_i, X_i, \boldsymbol{\beta})\, q(\theta_i|\boldsymbol{\Delta}) d\theta_i - \lambda \sum_j |\beta_j| \,, \tag{10}$$

where $\boldsymbol{\Delta}$ denotes the set of mean and variance of θ if they are estimable, and N_j denotes the number of individuals who answer item j. $L(\boldsymbol{Y}_i|\theta_i, \boldsymbol{\beta}, \boldsymbol{\Delta})$ is the likelihood function for person i, and $q(\theta_i)$ is the density of θ. A non-zero β_j for item j implies that this item exhibits uniform DIF for the focal/reference pair. λ is the tuning parameter that controls the magnitude of penalty.

Maximizing the marginal log-likelihood in Equation 10 is not computationally feasible. Instead, we use the expectation-maximization (EM) algorithm proceeding as follows. In the E-step, we construct the conditional expectation of the complete data log-likelihood with respect to θ. Suppose at the $(r + 1)$th EM cycle, we have,

$$Q^{r+1}(\boldsymbol{\beta}, \boldsymbol{\Delta}) = E_{\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{\Delta}^r, \boldsymbol{\beta}^r}\left(\log\left(L(\boldsymbol{\beta}, \boldsymbol{\Delta}|\boldsymbol{Y}, \boldsymbol{\theta})\right)\right) \equiv \sum_{j=1}^{J} Q_j^{r+1}(\beta_j) + Q^{r+1}(\boldsymbol{\Delta}) \tag{11}$$

where

$$Q_j^{r+1}(\beta_j) = \sum_{g=1}^{G}\sum_{i=1}^{N_{gj}} \int \log L(\beta_j|\boldsymbol{Y}_i, \theta, b_j) h(\theta|\boldsymbol{Y}_i, \boldsymbol{b}, \boldsymbol{\beta}^r) d\theta, \qquad (12)$$

where $h(\theta|\boldsymbol{Y}_i, \boldsymbol{b}, \boldsymbol{\beta}^r)$ is the posterior density of θ given the current estimate of $\boldsymbol{\beta}^r$ and other known information. G denotes total number of groups, and in our case G = 2, N_{gj} is the number of people in each group who answers item j. $Q^{r+1}(\boldsymbol{\Delta})$ is an integration of normal density with unknown mean μ_g and variance σ_g^2.

In the M-step, μ_g and σ_g^2 will be updated using the following closed forms.

$$\hat{\mu}_g = \frac{\sum_{m=1}^{M} n_{gjm} q_m}{N_{gj}},$$

$$\hat{\sigma}_g^2 = \frac{\sum_{m=1}^{M} n_{gjm} (q_m - \hat{\mu}_g)^2}{N_{gj}},$$

where $n_{gjm} = \sum_{i=1}^{N_{gj}} h(q_m|\boldsymbol{Y}_i, \boldsymbol{b}, \boldsymbol{\beta}^r)$ is the expected number of persons in group g and mth quadrature bin who answers item j. q_m is the mth quadrature along the θ scale. For the item DIF parameters, β_j, we can optimize it for different items separately, i.e.,

$$\beta_j = \max\{Q_j^{r+1}(\beta_j) - \lambda|\beta_j|\}, \qquad (13)$$

where $Q_j^{r+1}(\beta_j)$ defined in Equation 11 will be numerically approximated by $Q_j^{r+1}(\beta_j) = \sum_{g=1}^{G}\sum_{m=1}^{M} n_{gjm} \log L(\beta_j|\boldsymbol{Y}_i, q_m, b_j)$. Again, because this is a univariate optimization problem, the L-BGFS-B method available in the "optimize" function will be used.

To find an appropriate value of tuning parameter λ in Equation 12 and Equation 13, the Bayesian information criterion (BIC) is applied. We repeat parameters estimating process with different values of λ. Each λ value will give us an estimate of $(\boldsymbol{\beta}, \boldsymbol{\Delta})$. For the conditional likelihood method, the formula of BIC is given by

$$BIC = -2 \log L(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\theta}) + \|\boldsymbol{\beta}\|_0 \log N^* \qquad (14)$$

For the marginal likelihood method, we calculate a log-marginal likelihood for BIC that can be written as

$$BIC = -2\log\int P(\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{X},\boldsymbol{\beta})\, q(\boldsymbol{\theta}|\boldsymbol{\Delta}) d\boldsymbol{\theta} \; + \|\boldsymbol{\beta}\|_0 \log N^* \qquad (15)$$

In Equation 14, the first term is -2 times log-marginal likelihood of the observe data and the second term is the L_0 norm of $\boldsymbol{\beta}$ estimates times log sample size N^*. In CAT, different items tend to have different exposure rate, so there are many choices of N^*. Here we take N^*=min(N_j), which is the minimum of sample size in responses of all items. The model with the smallest BIC will be selected.

Simulation Study

Design

A simulation study was designed to evaluate the performance of the proposed two differential item function (DIF) detection methods in manipulated conditions imitating our real data example. The sample size was fixed to be 20,000, and item bank was fixed at 1000. True item difficulty, the *b* parameter, was generated from a standard normal distribution to capture a range of values. We assumed one reference and one focal group, with 50% of the sample in each group. Three DIF factors were manipulated: proportion of DIF items (1%, 5% and 10%), DIF magnitude, and impact. Three small proportions were considered because of the large item bank of CAT. For DIF magnitude, the *b* parameter for the studied items in the focal group were set to 0.25 (small DIF), 0.50 (medium DIF), or 1.00 (large DIF) higher than the reference group (e.g., Oshima, Raju, & Flowers, 1997; Suh & Cho, 2014). The trait variable for the focal and reference groups were generated as $\theta \sim N(\mu, 1)$. We also considered a no-impact and impact scenario. In the conditions of no impact, both focal and reference groups had $\mu = 0$. In the conditions where impact exists, the reference group still had $\mu = 0$, and the mean of θ for the focal group was set to 0.5. Altogether, there are 18 manipulated conditions. Each condition was simulated for 25 replications.[2] To simulate continuous responses y_{ij} for person i and item j from SRT, we used the inverse-Cumulative distribution function (CDF) method, where the CDF of y_{ij} follows $F(y) = \frac{\exp\left(y(\theta_i - b_j)\right) - 1}{\exp(\theta_i - b_j) - 1}$.

During CAT, the first five items were randomly selected, after $\hat{\theta}$ is obtained, either the match-b that criterion selects the next item whose b-parameter is closest to the current interim $\hat{\theta}$ was used or random selection was used. Random selection was included just as a reference baseline. Again, to imitate our real data example, variable-length design was considered, and the stopping rule was set as follows: the test stops either when the interim standard error of $\hat{\theta}$ is below 0.25 or when the test length reaches a maximum of 100.

Results

Table 1 displays the θ recovery results for the two item selection criteria under different DIF conditions as a quality control check from which several conclusions can be drawn. First, match-b item selection outperforms random selection by a noticeable margin in terms of mean bias and RMSE in fixed length CAT, as well as mean bias, RMSE and test length in variable-length CAT. However, unexpectedly, the correlation between true and estimated θ is higher from random selection compared to match-b selection. Given this, we plotted the true vs. estimated θs in Figure 1. As can be seen, although CAT with random item selection had a higher correlation (i.e., less coefficient bias), the variance of θ was underestimated such that the estimated traits are much more biased than the results of CAT with match-b item selection.

Table 1 results also showed that there were no appreciable differences in θ estimation

[2] In our preliminary check, we noticed that results start to stabilize even after 10 replications, hence 25 appears to be a reasonable choice.

between different DIF magnitude and DIF proportion conditions. Additionally, when there was impact, the bias of estimated θ became significantly larger especially in the match-b selection method. We know that IRT models provide the most information when the difficulty parameter is equal to the person and that is the reason why the match-b selection criterion is used in CAT design. When there was impact, we had $\theta \sim N(0.5, 1)$ for the focal group whereas item difficulty parameters were generated from $N(0, 1)$. So, for examinees in the focal group with a large θ, there were less informative items available in the item bank and as such, the measurements of θ were less accurate.

Figures 2-5 illustrate the average power (proportion of DIF items correctly detected) and Type I error rate for each condition (proportion of non-DIF items mistakenly flagged), with Figures 2-3 showing the results for the match-b CAT design, and Figures 4-5 displaying the results for the random selection CAT design. Within each figure, the three rows include results for the three different levels of DIF magnitude, and the two columns show power and Type I error, respectively.

As can be seen across the figures, for all three methods, increased DIF magnitude led to higher power. In addition, both Type I error and power appeared to be relatively stable with respect to the DIF prevalence, with power dropping slightly when the proportion of DIF items increases. This is understandable in that higher DIF prevalence may lead to slightly biased $\hat{\theta}$ that will in turn affect the methods that rely on $\hat{\theta}$ (i.e., CATSIB and conditional Lasso method). It can also be seen that the power for DIF detection in random selection CAT was better than for the match-b selection CAT, which is understandable because in random selection CAT, the response data we use to estimate an item's DIF includes a wider range of trait scores comparative to match-b selection CAT. Indeed, in a typical calibration scenario, heterogeneous sample would also yield more accurate item parameter estimation (e.g., Wang et al., 2018, JEM).

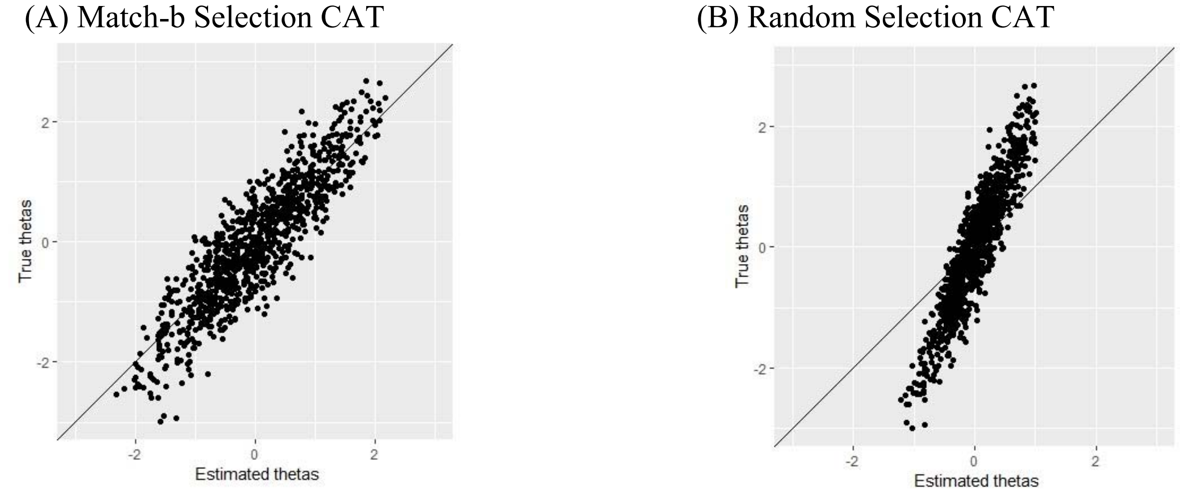
In addition, the two methods showed very different performance when impact exist. For the CATSIB, having impact seemed to yield higher power and slightly inflated Type I error. This was because θ estimate was biased when impact existed, as reflected by the large bias on those conditions. The regression approach in the SIBTEST, which was supposed to correct for bias, may not be as effective because $E_g[\theta]$ was also biased, as shown by large bias in the impact condition in Table 1. Note that under no impact condition, random selection led to higher power (with controlled Type I error) compared to match-b item selection. A possible reason is that as shown in Figure 1, $\hat{\theta}$ from random selection was much tighter. Hence assigning $\hat{\theta}$ to Q quadrature bins would yield more individuals in each bin. Hence $n_{R,q}$ and $n_{F,q}$ in the modified CATSIB equation (see Equation 6) would be larger for random selection, yielding smaller standard error and more powerful statistic.

For the conditional likelihood lasso method, the estimation of DIF parameter $\boldsymbol{\beta}$ was a simple maximization problem, and since both item difficulty parameter b and latent trait estimates $\hat{\theta}$ were known, we only needed the data of the focal group to estimate $\boldsymbol{\beta}$. We can see this method performed well except when impact was simulated in random selection CAT. In Table 1 we noticed that the bias of focal group trait estimates was high in the impact conditions in random selection CAT, and this might cause the worse performance of the conditional likelihood lasso method. On the other hand, the marginal likelihood lasso method estimates the latent trait distribution instead of using $\hat{\theta}$ from CAT. So, it worked stably well when impact exists in random selection design.

Although the Lasso with marginal likelihood method shows promising results, its power may be further improved because the current type I errors are very small in all conditions. That means, the information criterion can be further tuned. A drawback of the marginal likelihood method is the computing time. During the estimation, marginal likelihood needs to be calculated in every EM cycle. In the real data application, if match-b item selection CAT is used, conditional likelihood method would be a more economical choice.

Table 1 *Comparison of Different CAT Designs for θ Recovery (Reference Group / Focal Group)*

Impact	DIF size	DIF%	Match-b selection criterion			Random selection criterion		
			Mean bias	MSE	Correlation	Mean bias	MSE	Correlation
No Impact	+0.25	1%	-0.001/-0.002	0.227/0.227	0.880/0.880	-0.001/-0.001	0.418/0.419	0.911/0.910
		5%	-0.001/-0.001	0.228/0.227	0.879/0.880	-0.001/-0.001	0.419/0.419	0.911/0.910
		10%	-0.001/-0.001	0.228/0.229	0.879/0.879	-0.001/-0.001	0.419/0.419	0.911/0.910
	+0.5	1%	-0.002/-0.001	0.229/0.230	0.879/0.879	-0.001/-0.001	0.418/0.419	0.911/0.911
		5%	-0.001/-0.001	0.229/0.229	0.879/0.879	-0.001/-0.001	0.419/0.418	0.911/0.911
		10%	-0.001/-0.003	0.229/0.230	0.879/0.878	-0.001/-0.001	0.418/0.419	0.911/0.910
	+1	1%	-0.001/-0.008	0.228/0.223	0.879/0.880	-0.001/-0.004	0.419/0.418	0.911/0.911
		5%	-0.0003/-0.037	0.228/0.331	0.879/0.878	-0.001/0.002	0.419/0.419	0.912/0.911
		10%	-0.001/-0.009	0.228/0.330	0.878/0.878	-0.001/-0.0009	0.419/0.419	0.911/0.911
Impact +0.5	+0.25	1%	-0.0003/-0.127	0.227/0.244	0.880/0.879	-0.001/-0.314	0.420/0.519	0.911/0.910
		5%	7e-5/-0.136	0.228/0.244	0.879/0.879	-0.001/-0.317	0.419/0.522	0.911/0.911
		10%	-0.001/-0.145	0.228/0.245	0.880/0.879	-0.001/-0.322	0.419/0.523	0.911/0.910
	+0.5	1%	0.001/-0.129	0.227/0.247	0.880/0.879	-0.001/-0.314	0.419/522	0.912/0.911
		5%	-0.001/-0.126	0.228/0.247	0.879/0.879	-0.0009/-0.314	0.418/0.523	0.912/0.912
		10%	0.001/-0.128	0.227/0.247	0.879/0.879	-0.001/-0.337	0.418/0.523	0.911/0.911
	+1	1%	0.001/-0.160	0.228/0.248	0.879/0.878	0.001/-0.317	0.417/0.524	0.911/0.910
		5%	-0.001/-0.187	0.227/0.247	0.879/0.879	-0.001/-0.324	0.418/0.531	0.911/0.911
		10%	-0.001/-0.201	0.228/0.248	0.879/0.878	-0.001/-0.322	0.418/0.527	0.911/0.911


(A) Match-b Selection CAT

(B) Random Selection CAT

Figure 1 Scatter plots of true and estimated θ under two CAT cases

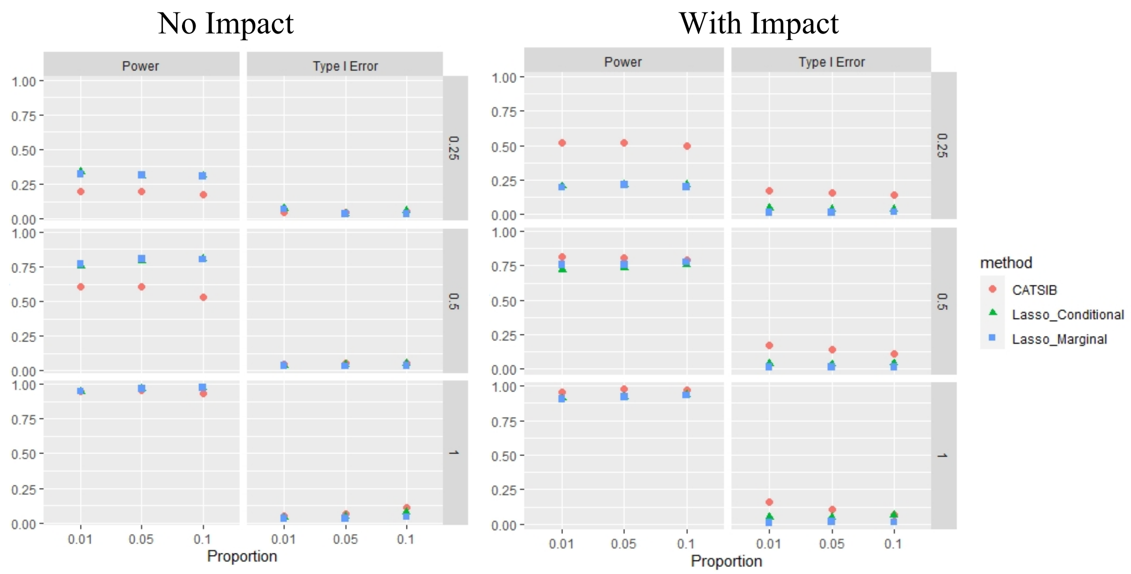


Figure 2 *Power and type I error of three methods in match-b selection CAT*

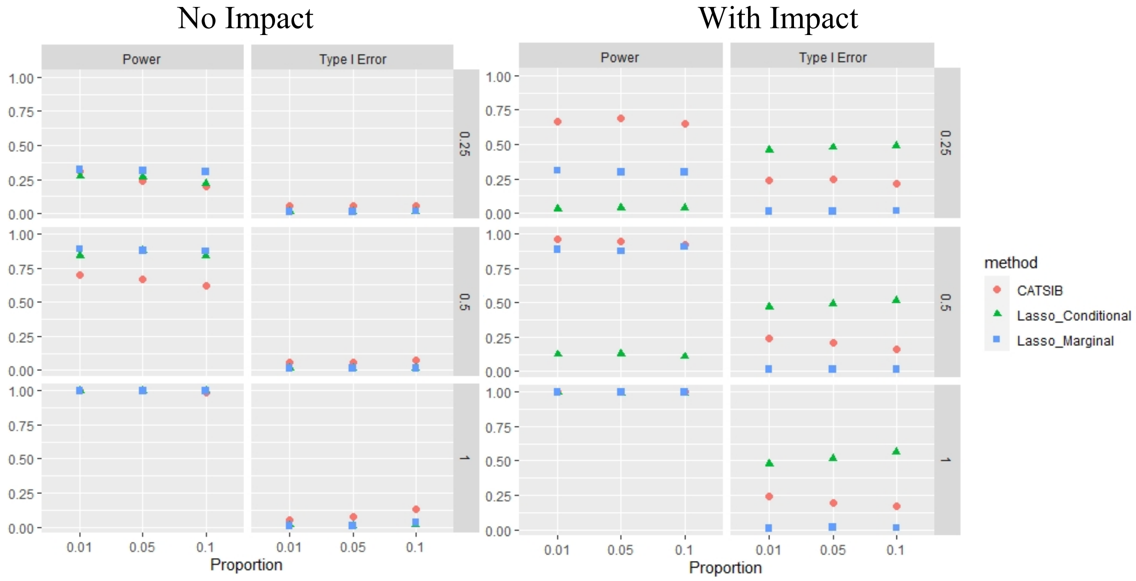


Figure 3 *Power and type I error of three methods in random selection CAT*

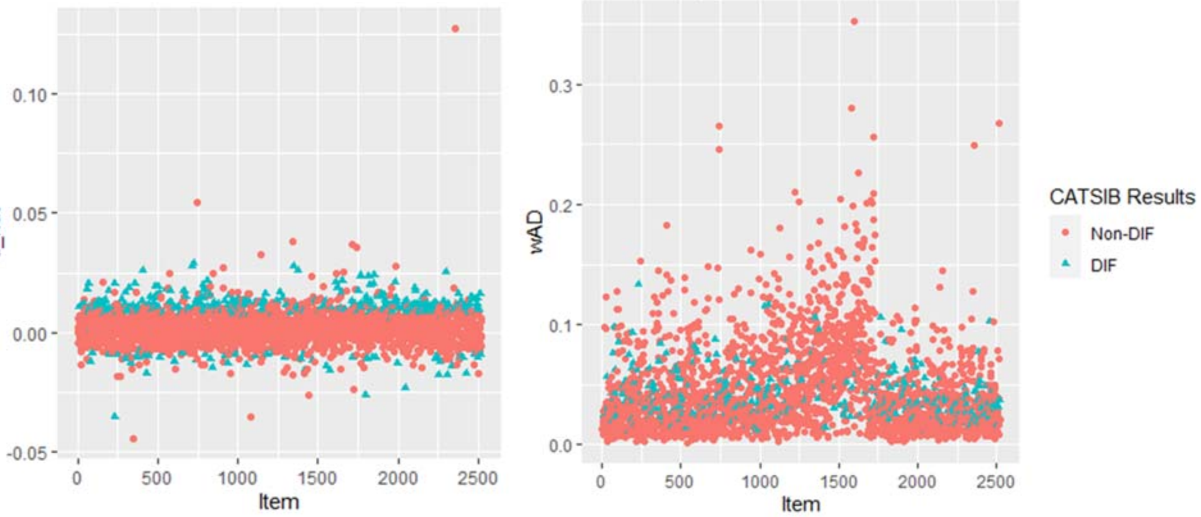
A Real Data Example

We obtained a real data set from a variable-length adaptive language assessment, which contains 110,260 individuals and each answered 6 speaking items. We focused only on speaking items because they were scored between 0 and 1 whereas the SRT model may not be suitable for other item types. There are 3,193 items in the bank, and hence the response matrix was quite sparse. To ensure adequate data per item, we used a relatively arbitrary cutoff of 20, that is, we removed 672 items that fewer than 20 individuals answered. Due to the data sparsity and large item bank, instead of using traditional marginal MLE for model calibration, the item difficulty parameters were obtained from large-scale machine-learning driven language models. They were provided in the data sets, and the mean difficulty is 5.21 and standard deviation is 2.03. Everyone's ability was also provided, and they were obtained by pooling information from other item types (such as sentence completion test, vocabulary test, etc.) to exploit the potential high correlation among different subdomains of language proficiency. However, as we will explain in detail below, such provided ability estimates (denoted as $\hat{\theta}^p$ where "p" stands for provided) can be used in the modified CATSIB method but not in the regularization method. We first report the modified CATSIB results.

The average $\hat{\theta}^p$ for the male and females are 5.029 and 5.057 respectively, hence we do not observe much impact existed in the data and therefore, with or without the regression correction in Equation 4 did not seem to make a difference. Figure 4a presents each item's $\hat{d}$ (i.e., Equation 5) with color-coded markers to denote detected DIF items vs. non-DIF items. Overall, 336 items were flagged to have DIF among 2,521 studied items. As shown, the identified DIF items tend to have more extreme $\hat{d}$'s than the non-DIF items. There are a few exceptions, which could be due to their large standard error computed from Equation 6 such that the B-statistic (from Equation 7) is no longer significantly different from 0. To further validate the DIF detection results, we compute the weighted absolute difference (wAD) between the empirical item characteristic curve from the focal and reference group, like the wABC statistics in Edelen et al. (2015). It is computed as

$$wAD = \int |P_F(\theta) - P_R(\theta)| f(\theta) d\theta \approx \sum_q^Q \left|\bar{P}_F(\theta_q) - \bar{P}_R(\theta_q)\right| w(\theta_q) \text{ (16)},$$

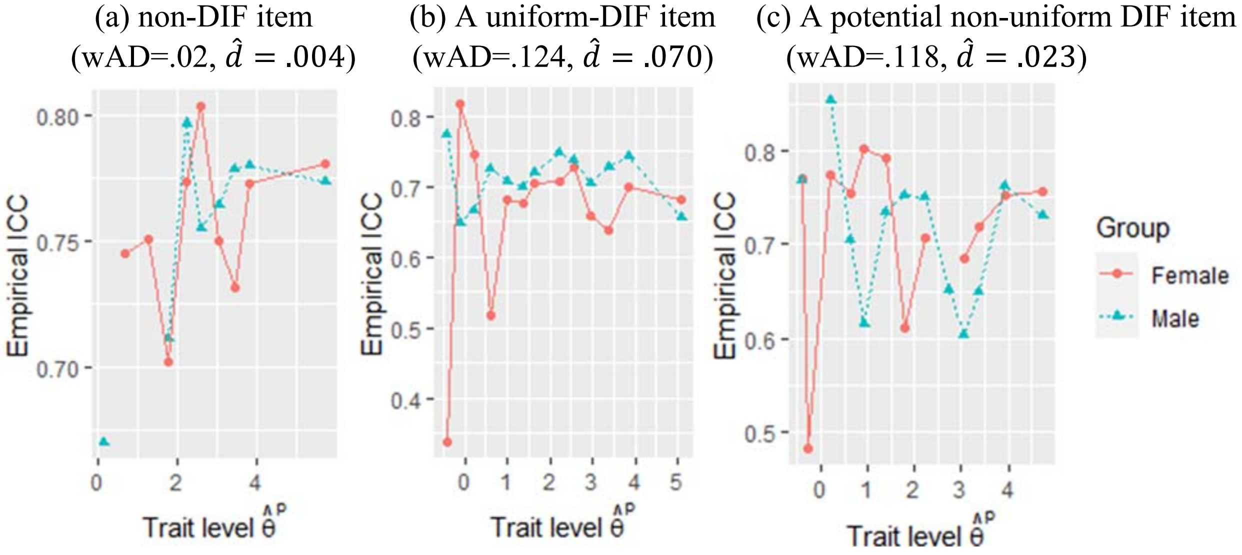
where $\bar{P}_F(\theta_q)$ and $\bar{P}_R(\theta_q)$ are the average scores people in the focal and reference group in the qth interval of $\hat{\theta}^p$. Without imposing any distributional assumptions on θ, we considered $f(\theta)$ as an empirical density of the θ distribution in the sample, i.e., from a pre-specified interval of -4 to 4 that covers all data in the sample, we use a grid of 0.2 and count the proportion of sample falling into each interval. As shown in Figure 4b, although the detected DIF items had relatively large wAD, there are quite many items with seemingly large wAD that were not flagged. This finding implies that some items may exhibit non-uniform DIF whereas the current modified CATSIB can only detect uniform DIF.


(a) $\hat{d}$

(b) wAD

Figure 4. Descriptive statistics ($\hat{d}$ and wAD) of DIF and non-DIF items differentiated by the modified CATSIB method

To further verify our detection results and conjecture about the existence of non-uniform DIF, we plot empirical item characteristic curve (ICC) per group for each item. Figure 5 presents the empirical ICC for three representative items, each of which has a sample size larger than 28 per group. Specifically, Figure 5a is a DIF-free item with a small wAD and a small $\hat{d}$; Figure 5b is an item that is flagged as DIF by the modified CATSIB method, with a large wAD and a large $\hat{d}$; Figure 7c is an item that is not flagged (i.e., small $\hat{d}$) but it still has a large wAD, and apparently the two ICCs cross.



(a) non-DIF item (wAD=.02, $\hat{d} = .004$) (b) A uniform-DIF item (wAD=.124, $\hat{d} = .070$) (c) A potential non-uniform DIF item (wAD=.118, $\hat{d} = .023$)

Figure 5. Empirical ICCs per group for three representative groups

With the absence of impact, we also expect the conditional and marginal regularization methods to perform similarly. However, note that the performance of the regularization method depends heavily on the model data fit. Hence, we first computed the weighted absolute difference between model predicted and empirical ICCs, denoted as $ICC\ wAD$. Given that the provided $\hat{\theta}^{p}$ was computed from an aggregate model that is different from SRT, it would be unfair to use $\hat{\theta}^{p}$ for evaluating the fit of SRT. Instead, we consider two ways of estimating estimate individual person's $\hat{\theta}$: (1) minimum cross-entropy (MCE; Maris, 2020)[3] and (2) MLE. In particular, $\hat{\theta}$ from the MCE method is used for the conditional Lasso method because it tends to be more accurate when test length is short, whereas $\hat{\theta}$ from MLE is used to compute wAD for the marginal Lasso method because we need to use a proper likelihood (i.e., SRT model likelihood, note that negative cross-entropy is not a likelihood) in the marginal Lasso method.

[3] Although the maximum likelihood estimation using the SRT likelihood in Equation 8 makes more sense theoretically, in a pilot check, we found that a minimum cross entropy method generates more precise individual $\hat{\theta}$ estimates due to small test length, i.e., each person was only administered 6 items in the real data.

After computing $\hat{\theta}$, we calculate $ICC\ wAD$ for each item as follows, and smaller value indicates better item-level model data fit.

$$ICC\ wAD_j = \int |P_M(\theta) - P_E(\theta)| f(\theta) d\theta \approx \sum_q^Q \left|\bar{P}_M(\theta_q) - \bar{P}_E(\theta_q)\right| w(\theta_q) \ (17),$$

where $f(\theta)$ is an empirical density function defined the same as in Equation 16. $\bar{P}_M(\theta_q)$ is the model predicted average score for people in the q^{th} interval, which is computed as follows for the marginal Lasso method (i.e., from SRT model)

$$\bar{P}_M(\theta_q) = \int_0^1 yP(y|\theta_q)dy = \frac{\left(1 - \frac{1}{\theta_q - b_j}\right)\exp(\theta_q - b_j) + \frac{1}{\theta_q - b_j}}{\exp(\theta_q - b_j) - 1}$$

where $P(y|\theta_q)$ is the density function defined in Equation 1. $\bar{P}_E(\theta_q)$ are the empirical, average scores people in the q^{th} interval received. $\bar{P}_M(\theta_q)$ is computed directly from Rasch model, i.e., $\frac{1}{1+\exp(-(\theta_q - b_j))}$, for the conditional Lasso method.

We first conducted the conditional Lasso method conditioning on the MCE estimated $\hat{\theta}$. Note that in this method, only data from focal group contributes to the estimation of $\boldsymbol{\beta}_j$ (see Equation 8). Hence, treating either male and female as focal group may or may not yield different results. To exercise due diligence, we ran both and found that the overlap of flagged items is as high as .79. This is interesting as it implies that these flagged items are the items that tend to not fit well with the ICCs defined by the Rasch model, i.e., the DIF detection is contaminated by the lack of item fit. Indeed, Figure 6 presents the $ICC\ wAD_j$ results for both the detected DIF and non-DIF items. As shown, the flagged DIF items tend to have larger $ICC\ wAD_j$.

Taking female as the focal group as in the CATSIB method, 166 items were flagged as DIF. However, the overlap of detection between the conditional Lasso and the modified CATSIB method is only .127. Like Figure 4, Figure 7[4] presents the $\hat{d}$ and wAD from those DIF and non-DIF items, and as shown, those flagged items are not necessarily the ones with larger $\hat{d}$ or large wAD, indicating that the DIF detection in this case is confounded with detection of lack of fit.

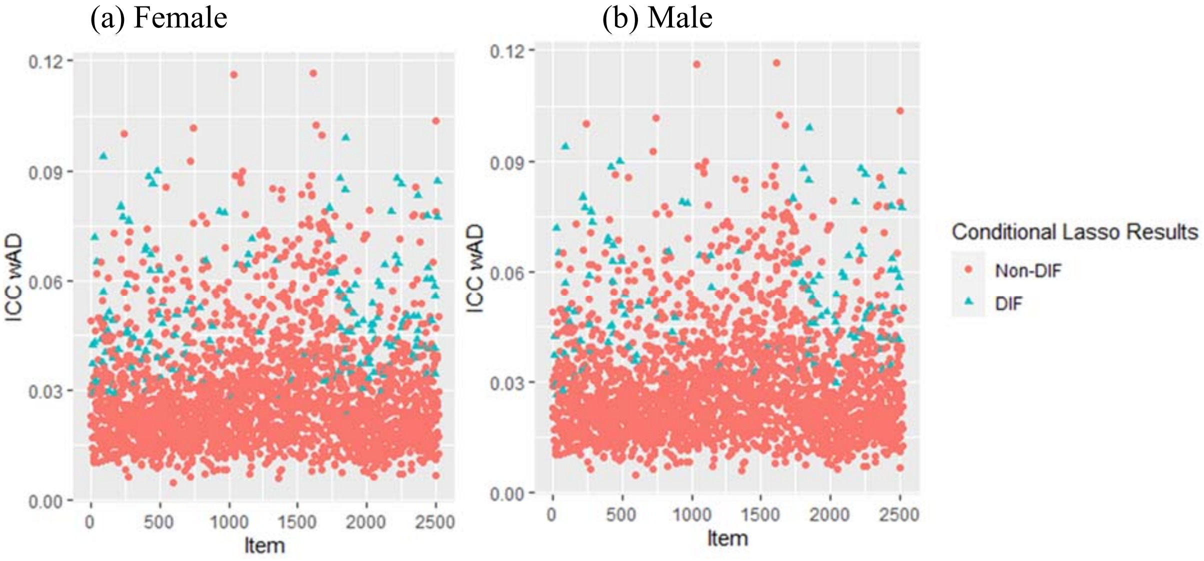

(a) Female (b) Male

Figure 6. $ICC\ wAD_j$ for DIF and non-DIF items (detected by the conditional Lasso method) using female (left) and male (right) as focal group.

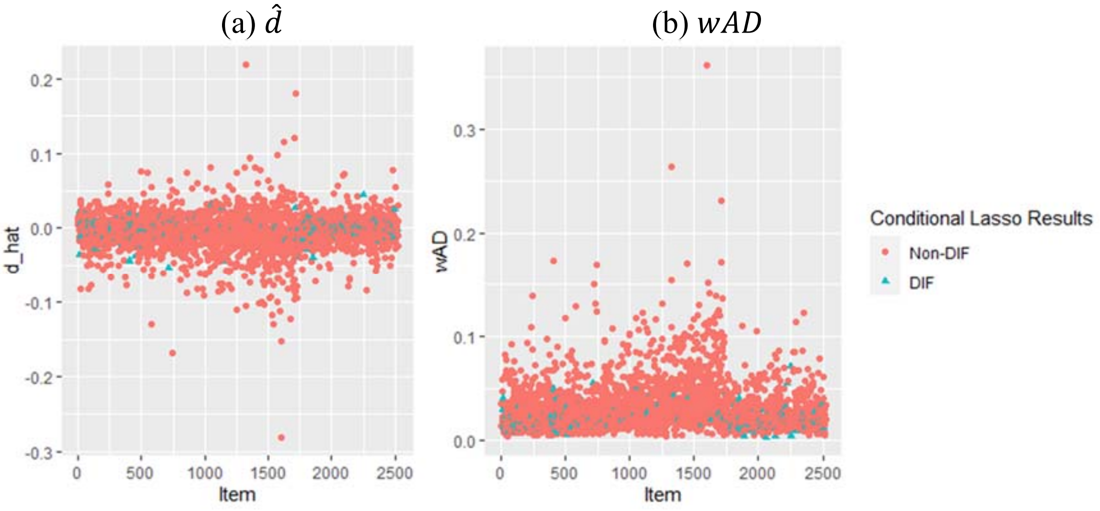

(a) $\hat{d}$ (b) wAD

Figure 7. Descriptive statistics ($\hat{d}$ and wAD) of DIF and non-DIF items differentiated by the conditional Lasso method

For the marginal Lasso method, although data from both focal and reference group were

[4] In Figure 7, we used $\hat{\theta}$ from MCE to compute the descriptive statistics, whereas $\hat{\theta}^P$ was used in Figure 4.

used for estimation, we still tried both ways of treating female and male as focal groups respectively and found the overlap of detected DIF items is as high as .974. Further, we used female group as the focal group as before, 1,849 items were flagged as DIF. Interesting, 158 out of 166 DIF items detected by the conditional Lasso method were flagged again by this method, although the marginal Lasso method flagged many more items. Figure 8 presents the $\hat{d}$, wAD, and $ICC\ wAD$ computed based on SRT MLE of $\hat{\theta}$. As shown, neither $\hat{d}$ nor wAD appeared to be large for the flagged DIF items, but surprisingly, $ICC\ wAD$ tended to separate DIF and non-DIF items well. This reinforces our conjecture that the detected "DIF" items are those that show lack of fit.

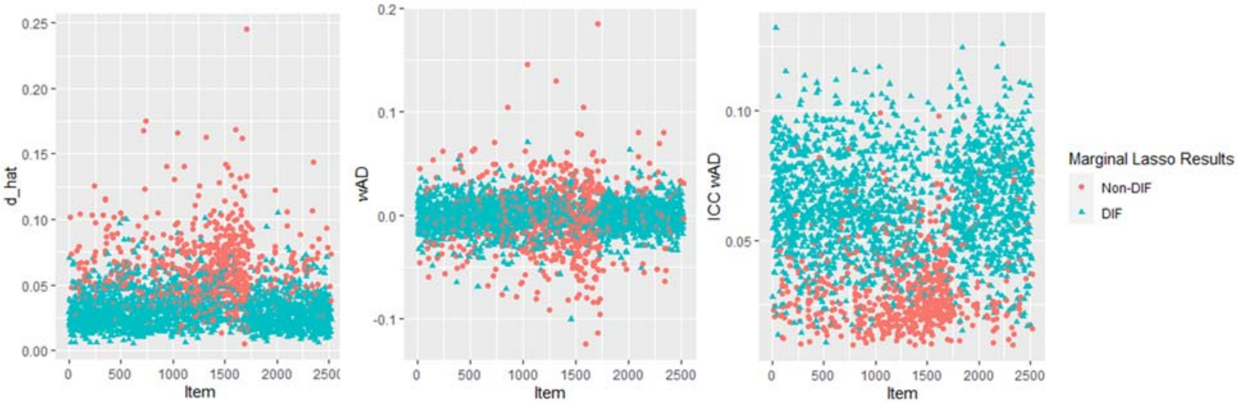


Figure 8. Descriptive statistics ($\hat{d}$ and wAD and ICC wAD) of DIF and non-DIF items differentiated by the marginal Lasso method

Discussion

In this paper, we extended the CATSIB and regularization methods to detect DIF in continuous, severely sparse CAT. Indeed, the rapidly expanding role of machine learning and artificial intelligence technologies in educational and psychological measurement spur a new genre of assessment which uses "machine learning" techniques to automatically generate items. Such a high-quality, large item pool greatly enables personalized assessment via state-of-the-art variable length CAT. However, there is no empirical research on DIF detection in this new, yet booming scenario. Our simulation studies showed that both methods hold great promise in detecting DIF for such data type. Specifically, the conditional Lasso method is recommended if the model fits the data, whereas the modified CATSIB is a viable model-free alternative otherwise. The marginal Lasso method, although is statistically optimal in theory, it may be computationally too expensive to apply in practice.

The real data example sheds further light on the application potential of the two methods. That is, the modified CATSIB method appears to be effective to detect items with uniform DIF, although certain effect size measure cutoff can be used in practice to distinguish statistically significant DIF from practically meaningful DIF, such as the standardized mean difference in item scores across groups (Dorans & Kulick, 1986; Zwick, Donoghue, & Grima, 1993). The regularization methods, on the other hand, rely on the adequate model data fit. It appears that those flagged DIF items also tend to be items shown misfit, implying that item misfit may confound with DIF detection. This is especially true for conditional Lasso method: because DIF is reflected by non-zero β_j in Equation 8, and given known θ, only data from focal group contributes to the estimates of β_j. Hence non-zero β_j could also mean the current SRT model does not fit the data from focal group well. In fact, if we use the true focal group as the focal group and true reference group as the reference group, we will observe non-zero β_j's for DIF items, and if we reverse the role of focal and reference group, we will observe zero β_j's instead. However, in our analysis, when we used both female and male group as focal group in the two separate analyses, we noted that almost the same set of DIF items were detected. This reinforces that it may be the lack of fit, rather than DIF, that contributes to the non-zero β_j's. In this case, results from the modified CATSIB method should hold higher credence.

The current study can be expanded in a few directions. First, since we considered a "Rasch" version of a continuous response model (i.e., the SRT model) throughout the study, our focus was on the uniform DIF. However, the modified CATSIB method can be further extended to detect non-uniform DIF, such as in Chalmers (2018). The primary idea is to find θ_c at which $P_R(\theta_c) - P_F(\theta_c) = 0$, i.e., the two ICCs cross, and then update Equation 3 as follows,

$$d = \int_{\theta<\theta_c} (P_R(\theta) - P_F(\theta)) f(\theta)\, d\theta + \int_{\theta\geq\theta_c} (P_F(\theta) - P_R(\theta)) f(\theta)\, d\theta \text{ (18).}$$

Chalmers (2018) derived a statistic based on Equation 18 which follows a chi-squared distribution. Second, other continuous item response model may be considered in addition to SRT, such as Samejima (1973) which assumes the probability of an observed response larger than or equal to a constant follows a normal cumulative distribution function (Zopluoglu, 2013). Samejima (1973)'s model can be viewed as a limiting form of the graded response model, whereas the SRT model considered herein was derived from a sufficient statistic scoring rule.

Reference

Bauer, D. J., Belzak, W. C., & 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.

Belzak, W., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*.

Cardwell, R., LaFlair, G. T., and Settles, B. (2022). *Duolingo English Test: Technical Manual.* Available Online at: http://duolingo-papers.s3.amazonaws.com/other/det-technical-manual-current.pdf.

Chalmers, R. P. (2018). Improving the crossing-SIBTEST statistic for detecting non-uniform DIF. *Psychometrika, 83*(2), 376-386.

Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*(3), 333-353.

Dorans, N. J., & 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.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software, 33*(1), 1.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. *Test validity*, 129-145.

Khodadady, E. (2014). Construct validity of C-Tests: A factorial approach. Journal of Language. Teaching and Research, 5(6), 1353- 1362

Lei, P. W., Chen, S. Y., & Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement, 43*(3), 245-264.

Magis, D., Tuerlinckx, F., & De Boeck, P. (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics, 40*(2), 111-135.

Maris, G. (2020). *The Duolingo English Test: Psychometric considerations*. Duolingo Research Report DRR-20-02, 1-11.

Maris, G., & Van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika, 77*(4), 615-633.

Nandakumar, R., & Roussos, L. (1997). Validation of CATSIB To Investigate DIF of CAT Data.

Nandakumar, R., & Roussos, L. (2004). Evaluation of the CATSIB DIF procedure in a pretest setting. *Journal of Educational and Behavioral Statistics, 29*(2), 177-199.

Oshima, T., Raju, N. S., & Flowers, C. P. (1997). Development and Demonstration of Multidimensional IRT-Based Internal Measures of Differential Functioning of ltems and Tests. *Journal of Educational Measurement, 34*(3), 253-272.

Penfield, R. D., & Camilli, G. (2006). 5 Differential Item Functioning and Item Bias. *Handbook of statistics, 26*, 125-167.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38*(2), 203-219.

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

Suh, Y., & Cho, S.-J. (2014). Chi-square difference tests for detecting differential functioning in a multidimensional IRT model: A Monte Carlo study. *Applied psychological measurement, 38*(5), 359-375.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the royal statistical society. Series B (methodological)*, 267-288.

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika, 80*(1), 21-43.

Wang, C., Chang, H.-H., & Boughton, K. A. (2013). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied psychological measurement, 37*(2), 99-122.

Wang, C., Chen, P., & Jiang, S. (2020). Item calibration methods with multiple subscale multistage testing. *Journal of Educational Measurement, 57*(1), 3-28.

Wang, C., Weiss, D. J., & Shang, Z. (2018). Variable-Length Stopping Rules for Multidimensional Computerized Adaptive Testing. *Psychometrika*, 1-23.

Wang, C., Zhu, R., & Xu, G. (2021). Using Lasso and Adaptive Lasso to Identify DIF in Multidimensional 2PL Models. *Multivariate behavioral research*, 1-21.

Wang, T., & Zeng, L. (1998). Item parameter estimation for a continuous response model using an EM algorithm. *Applied psychological measurement, 22*(4), 333-344.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied psychological measurement, 35*(5), 339-361.

Zopluoglu, C. (2013). A comparison of two estimation algorithms for Samejima's continuous IRT model. *Behavior research methods, 45*(1), 54-64.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233-251.

Zwick, R., Thayer, D. T., & Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied psychological measurement, 18*(2), 121-140.