Article

Educational and Psychological
Measurement
2016, Vol. 76(1) 22–42
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164415579488
epm.sagepub.com
SAGE

Item Response Theory With Covariates (IRT-C): Assessing Item Recovery and Differential Item Functioning for the Three-Parameter Logistic Model

Louis Tay[1], Qiming Huang[1], and Jeroen K. Vermunt[2]

Abstract

In large-scale testing, the use of multigroup approaches is limited for assessing differential item functioning (DIF) across multiple variables as DIF is examined for each variable separately. In contrast, the item response theory with covariate (IRT-C) procedure can be used to examine DIF across multiple variables (covariates) simultaneously. To assess the utility of the IRT-C procedure, we conducted a simulation study. Using SAT data for realistic parameters, uniform DIF on three covariates were simulated: gender (dichotomous), race/ethnicity (categorical), and income (continuous). Simulations were conducted across several conditions: two test lengths (14 items, 21 items), four sample sizes (5,000, 10,000, 20,000, 40,000), and two DIF effect sizes (medium, large). It was found that the IRT-C procedure could accurately recover the latent means and the three-parameter logistic model parameters well with a substantial sample size of 20,000. There was good control of Type I error rates to the nominal rates across the sample sizes. Good power to detect DIF across all covariates ($>.80$) was observed when the sample size was 20,000 for large DIF effect size and 40,000 for medium DIF effect size. Practical implications for the use of the IRT-C procedure are discussed.

[1]Purdue University, West Lafayette, IN, USA
[2]Tilburg University, Tilburg, Netherlands

Corresponding Author:
Louis Tay, Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907, USA.
Email: stay@purdue.edu

Keywords

differential item functioning, item response theory, simulation, covariates

The assessment of differential item functioning (DIF) in large-scale testing relies on the multigroup approach to assessing DIF. However, with the multigroup approach, it is difficult to assess DIF across multiple variables (e.g., age and gender) simultaneously as DIF is commonly assessed on each variable separately. This limits the extent we can isolate the proximal cause(s) of DIF (e.g., Robert, Lee, & Chan, 2006); that is, because we examine DIF on each variable (e.g., age, gender) in turn, it is difficult to know whether DIF is attributable to the different demographic, attitudinal, or cultural dimensions. For example, the same item may display DIF on age and gender, and it may be difficult to know whether DIF is attributable to one of both variables. Another potential problem with the multigroup approach is that it requires the nominalizing of variables (e.g., age → age groups; e.g., Kim, Cohen, & Park, 1995) and categorization leads to a loss of information and power to detect DIF. By estimating DIF across multiple variables simultaneously, we can more accurately determine the latent trait scores with respect to different variables. For instance, we can determine the impact of age on trait scores controlling for DIF of age, while simultaneously controlling for DIF and impact of other variables such as gender, socioeconomic status, and race/ethnicity.

Recent research has examined the use of a factor analytic framework, known as the multiple indicators multiple cause (MIMIC) model, for assessing DIF across multiple variables without a need to nominalize variables (Woods, 2009a, 2009b; Woods & Grimm, 2011; Woods, Oltmanns, & Turkheimer, 2009). However, it is not suitable for test data because guessing is not modeled in a factor analytic framework. There have been developments in the item response theory (IRT) counterpart of the MIMIC model, called the IRT with covariates (IRT-C) model (Tay, Newman, & Vermunt, 2011), which allows for the modeling of guessing parameters. Nevertheless, past research examining the utility (e.g., power and Type I error rate) of the IRT-C approach for assessing DIF has been limited to only the two-parameter logistic model (2PLM) and one covariate (Tay, Vermunt, & Wang, 2013). In view of this, the primary goal of this article is to assess the utility (e.g., power and Type I error rate) of the IRT-C for assessing DIF across multiple covariates in a testing context where a three-parameter logistic model (3PLM) is used. Practically, we seek to determine the conditions (viz. sample size and test length) for which the IRT-C can be used to assess DIF across multiple covariates in tests.

IRT-C Model

The IRT-C model is used to model item endorsements as a function of the latent trait while accounting for DIF from multiple covariates simultaneously (Tay et al., 2011). We first describe the 3PLM and later describe how covariates are modeled.

A 3PLM can be used to describe the relationship between item endorsement y_{ij} and the latent trait level θ_j, where subscript $j = 1, \ldots, J$ indexes individuals and subscript $i = 1, \ldots, I$ indexes items. The probability of item endorsement is given as

$$P(y_{ij}|\theta_j) = c_i + \frac{1 - c_i}{1 + \exp\left(-\left[a_i\theta_j + b_i\right]\right)}, \tag{1}$$

where a_i, b_i, and c_i, represent the item discrimination, item location, and pseudo-guessing parameter, respectively.

DIF occurs when the probability of item endorsement differs as a function of an observed characteristic z_j, or covariate, given the same latent trait level θ_j. As such, the probability of endorsement is given as

$$P(y_{ij}|\theta_j, z_j) = c_i + \frac{1 - c_i}{1 + \exp\left(-\left[a_i\theta_j + b_i + c_i z_j + d_i z_j \theta_j\right]\right)}, \tag{2}$$

where c_i and d_i represents the coefficients for the direct and indirect effects of the observed characteristic z_j, respectively. The direct effect corresponds to uniform DIF and the indirect effect corresponds to nonuniform DIF. By testing for the significance of c_i and d_i using the Wald χ^2 statistic, we can determine whether uniform and nonuniform DIF, respectively, is significant. By extension, it is possible to extend Equation (2) to a vector of observed characteristics $\mathbf{z_j}$, or covariates, to test for DIF on multiple covariates at once.

In addition, the distributions of the latent traits across the observed characteristics are modeled

$$f(\theta_j|z_j) \sim N(\mu_z, \sigma_z^2), \tag{3}$$

where μ_z and σ_z^2 refer to the latent mean and variance corresponding to the observed characteristic z. The reference group latent trait mean and variance are fixed to 0 and 1, respectively, while the focal group latent trait parameters are freely estimated for model identification.

IRT-C Procedure for Uniform DIF

The proposed IRT-C procedure (Tay et al., 2011) starts from a fully constrained baseline model shown in Equation (1) where no DIF is modeled; that is, all c_i and d_i parameters in Equation (2) are constrained to zero. Starting from a fully constrained baseline model instead of an unconstrained model overcomes the problem of model nonidentification when all c_i and d_i parameters are estimated. Furthermore, there is no need to know anchor items a priori, which is often difficult when seeking to assess DIF across multiple covariates. For the purposes of this study, we seek to only examine the presence of uniform DIF and so present only the procedure associated with that.

The steps to identifying uniform DIF across multiple covariates are as follows:

1. For the current model, the possible presence of DIF is flagged by the bivariate residual (BVR) between the item and the covariate. The BVR is a standardized residual component between the item and covariate not accounted for by the model. It is analogous to a modification index and indicates the extent there is local dependency between the item and the covariate (likely as a result of DIF). The item–covariate pair that has the largest BVR will be flagged.
2. In this step, the flagged item–covariate pair will have the uniform DIF parameter c_i modeled and tested for significance using the Wald test. If DIF is not significant at a nominal rate of .05, the procedure ends and the previous model would be chosen as the final model. Otherwise, the procedure continues.
3. In this step, if the *c* parameter is significant, a model is specified with the significant *c* parameter shown in Equation (2). This less constrained model replaces the more constrained initial model in Step 1. All three steps (Steps 1, 2, and 3) are repeated iteratively until the item which has the highest BVR does not have significant DIF.

Additional stopping rules have also been proposed where different information criteria such as the Bayes Information Criterion (BIC) are compared between the more constrained (no DIF specified for target item) model and less constrained (DIF specified for target item) model; the procedure stops when the less constrained model is less parsimonious (higher information criterion) regardless of whether the *c* value is significant.

A past simulation study of the IRT-C 2PLM procedure for assessing DIF found good control of Type I error rates close to the nominal rate of .05 (Tay et al., 2013). Furthermore, it showed equivalent or better power to detect DIF as compared with the Mantel–Haenszel procedure and the MIMIC procedure where anchor items are assumed to be known. One advantage of this proposed IRT-C procedure is that it does not require the use of known anchor items or invariant items (e.g., Allalouf, Hambleton, & Sireci, 1999; Stark, Chernyshenko, & Drasgow, 2006), as these are often not known a priori in practice.

Current Study

In this study, we used the scholastic aptitude test (SAT) data for realistic parameters and simulated uniform DIF across three different covariates—gender (a dichotomous variable), race/ethnicity (a multicategory variable), and household income (a continuous variable). We examined the Type I error rates and power for detecting uniform DIF with the IRT-C approach using the two different stopping rules (with and without the BIC criterion). In addition, we also evaluate the accuracy of the estimated 3PLM item parameters and estimated latent means across the covariates.

Table 1. Percentages of Examinees by Sex and Race/Ethnicity.

%	White	Black	Hispanic	Asian
Male	29.94	5.34	6.01	4.54
Female	34.15	6.72	8.23	5.08

Note. Black = Black or African American; Hispanic = Mexican or Mexican American, Puerto Rican, Other Hispanic, Latino, or Latin American; Asian = Asian, Asian American, or Pacific Islander.

Method

We determine the extent to which two different IRT-C procedures can be used to assess uniform DIF on multiple covariates across two test lengths (Nitems = 14, 21), four different sample sizes (N = 5,000, 10,000, 20,000, 40,000), and two DIF effect sizes (DIF size = medium, large). For each of the 16 conditions, a total of 400 replications were undertaken. We used only a smaller number of simulation conditions as the proposed iterative IRT-C procedure took a substantial time to complete a single replication: 2 to 4 hours for a sample size of 5,000 and 22 to 30 hours for a sample size of 40,000. As such, the simulations took more than 64,000 computer hours to complete. In addition, in our preliminary examinations, longer test lengths (e.g., 35 items) took up too much memory to estimate (2 GB as implemented in the current software), and sample sizes that were smaller than 5,000 were not well-estimated with many replications not reaching convergence.

Data Generation

Covariates. In order to simulate relationships between the external covariates and the latent trait that have high fidelity to test data, we use descriptive statistics from a random sample of 100,000 examinees taking the 2008 SAT to predict the standardized SAT Math scores and using these predicted scores as simulated ability estimates. In this study, we used standardized SAT Math[1] scores as a proxy for the IRT latent trait because IRT scores have very high correlations with the observed scores. Table 1 displays the percentages of individuals for gender (two categories: Male, Female) and race/ethnicity (four categories: White, Black, Hispanic, Asian). To predict the standardized SAT Math scores, an additive regression model was fit to the data with predictor variables of average household income, gender, and race/ethnicity. Gender and race/ethnicity were dummy coded so that the reference group was White Male with average household income,

$$\text{SAT Math Score} = \beta_0 + \beta_1(\mathit{Female}) + \beta_2(\mathit{Black}) + \beta_3(\mathit{Hispanic}) + \beta_4(\mathit{Asian}) + \beta_5(\mathit{ZIncome}) + e,$$

where *ZIncome* and e represent the standardized household income and error term, respectively. The multiple R^2 value was 0.13 and the estimated beta coefficients are shown in Table 2.

Table 2. Results of Addition Regression Model Predicting SAT Math Scores.

	Estimated beta weights	Standard error
Intercept	0.1217	0.0060
Male	—	—
Female	0.0210	0.0070
White	—	—
Black	−0.6398	0.0120
Hispanic	−0.3568	0.0113
Asian	−0.0547	0.0130
ZIncome	0.2200	0.0040

Note. Black = Black or African American; Hispanic = Mexican or Mexican American, Puerto Rican, Other Hispanic, Latino, or Latin American; Asian = Asian, Asian American, or Pacific Islander.

Theta Distributions. To simulate a theta distribution that conforms to the estimated additive model, random normal distributions were simulated for each of the eight cells (gender × race/ethnicity) shown in Table 1. All the variances of the normal distribution were set at 1.00, but the mean of the random normal distributions was dependent on the beta coefficients, where

$$\theta^{*}\text{mean} = \hat{\beta}_0 + \hat{\beta}_1(Female) + \hat{\beta}_2(Black) + \hat{\beta}_3(Hispanic) + \hat{\beta}_4(Asian). \tag{5}$$

The standardized household income effect was simulated by (a) assigning a value from a normal distribution with the mean given in Equation (2) to each simulee, and (b) using the formula $\theta = \theta^{*} + \hat{\beta}_5(ZIncome)$, where θ^{*} is the initial theta value drawn from a normal distribution for each cell and θ is the final simulated theta value.

The number of simulees in each cell was based on the proportions in Table 1 multiplied by the total number of simulees. We note that this may not result in an exact N value because of rounding to a whole number within each cell.

Item Parameters. For a traditional 3PLM given by

$$P(y_{ji}|\theta_j) = c_i + \frac{1 - c_i}{1 + \exp(-1.702a_i^{*}[\theta_j - b_i^{*}])}, \tag{6}$$

item discriminations a_i^{*} were sampled from a truncated normal ($Mean$ = 1.2, SD = 0.3) with lowest and highest possible values set to 0.5 and 1.7, respectively; b_i^{*} was sampled from a uniform (−2, 2) distribution; c_i was sampled from a logit-normal (−1.1, 0.5) distribution. Final generated item parameters a_i and b_i were obtained by the transformations

$$a_i = 1.702 \times a_i^{*}$$

and

$$b_i = -1.702 \times a_i^* \times b_i^*.$$

The implemented 3PLM is

$$P(y_{ji}|\theta_j) = c_i + \frac{1 - c_i}{1 + \exp(-[a_i\theta_j + b_i])}. \tag{7}$$

Differential Item Functioning Across Multiple Covariates. In IRT-C, uniform DIF between males and females can be described by

$$P(y_{ij}|\theta_j, gender_j) = c_i + \frac{1 - c_i}{1 + \exp\left(-\left[a_i\theta_j + b_i + c_i gender_j\right]\right)}, \tag{8}$$

where the probability of item responding depends not only on θ_j but also gender, which is dummy coded (male = 0; female = 1). By extension, the equation can be expanded to include the other two covariates—race/ethnicity and household income. With the reference group set to White males with average household income, we simulate DIF on the first seven items shown through the design matrix that specified the DIF coefficients ()

$$\begin{array}{c} i=1 \\ i=2 \\ . \\ . \\ . \\ . \\ i=7 \end{array} \begin{bmatrix} .40 & 0 & 0 & 0 & 0 \\ .40 & 0 & 0 & 0 & -.20 \\ .40 & .40 & 0 & 0 & 0 \\ 0 & .40 & 0 & -.40 & 0 \\ 0 & .40 & 0 & -.40 & -.20 \\ 0 & 0 & 0 & 0 & -.20 \\ .40 & .40 & 0 & -.40 & -.20 \end{bmatrix} \begin{bmatrix} Female \\ Black \\ Hispanic \\ Asian \\ ZIncome \end{bmatrix}.$$

It is important to note that this matrix is used for the purposes of simulation and does not reflect how the test is biased against or for different groups of individuals. For moderate DIF, we specified uniform DIF of the magnitude of .40 on Items 1, 2, 3, and 7 biased in favor of females. Items 3, 4, 5, and 7 are biased in favor of Blacks, whereas Items 4, 5, and 7 are biased against Asians. Finally, Items 2, 5, 6, and 7 are biased against those with high household income. The design matrix was created to determine the extent the IRT-C procedure can detect DIF when it occurs only on a single covariate, or on multiple covariates. For large DIF, we specified uniform DIF of the magnitude .60 in place of .40, and .30 in place of .20.

Test Length. We simulated test lengths of 14 and 21 items where the first 7 items have DIF of the form specified in the DIF matrix but the remaining items do not have DIF. As such, our simulations focus on the boundary conditions with tests that have a large proportion of DIF items (50%) and a moderately large proportion (30%) of DIF items, respectively. We expect that even with a test where 50% of items have DIF on one of the covariates, we should still be able to identify DIF because not all the items have DIF on all the covariates.

Responses. Dichotomous responses were generated using Equation (4) by (a) drawing a uniform [0, 1] number (U) and (b) comparing it to $P(y_{ji}|\theta_j)$: if $P(y_{ji}|\theta_j) > U$, the simulated response = 1, otherwise the simulated response was 0.

Estimation

The software Latent GOLD 4.5 (Vermunt & Magidson, 2008) was used to conduct the estimation. The 3PLM was estimated of the item responses shown in Equation (4) and DIF was estimated based on Equation (6). Latent mean level differences due to external covariates was estimated with the equation:

$$\theta = b_0 + b_1(Female) + b_2(Black) + b_3(Hispanic) + b_4(Asian) \\ + b_5(ZIncome) + e.$$

Maximum likelihood estimation was used with a total of 30 quadrature nodes.

We note that Wald test implemented in Latent GOLD 4.5 uses improved estimation of the asymptotic sampling variance–covariance matrix resulting in better controlled Type I error rates (Langer, 2008; Woods, Cai, & Wang, 2012).

Criteria

For all the criteria, we report the average values across the 400 replications for the final estimated IRT-C model.

Covariate Coefficients. The root mean squared error (RMSE) was computed between the estimated covariate coefficients in Equation (7) and the simulated covariate coefficients.

3PLM Item Parameters. Two criteria were used to assess the accuracy of item parameter estimation. First, we examined the correlation between the estimated item parameters and simulated item parameters. Second, the RMSE was computed for the estimated item parameters.

Differential Item Functioning. First, Type I error rates and power were computed for each covariate separately. Type I error rates were computed as the proportion of non-DIF items that were flagged as having DIF. Power was computed as the proportion of DIF items that were flagged for DIF.

Results

Estimated Covariate Coefficients

The estimated covariate coefficients were similar across the DIF effect size conditions and we only present the medium effect size results here (see Appendix for large DIF effect size results). As shown in Table 3, the estimated covariate coefficients

Table 3. RMSE of Estimated and Simulated Covariate Coefficients.

		BIC stopping rule (RMSE)					No BIC stopping rule (RMSE)				
Test length	Sample size	Female	Black	Hispanic	Asian	ZIncome	Female	Black	Hispanic	Asian	ZIncome
Nitem = 14	*N* = 5,000	0.046	0.066	0.040	0.063	0.082	0.037	0.060	0.042	0.056	0.079
	N = 10,000	0.034	0.059	0.028	0.050	0.080	0.022	0.042	0.030	0.040	0.075
	N = 20,000	0.019	0.042	0.021	0.034	0.077	0.016	0.027	0.020	0.027	0.074
	N = 40,000	0.017	0.029	0.015	0.023	0.076	0.011	0.018	0.015	0.018	0.074
Nitem = 21	*N* = 5,000	0.034	0.052	0.039	0.050	0.079	0.029	0.048	0.037	0.048	0.075
	N = 10,000	0.023	0.043	0.025	0.039	0.077	0.020	0.032	0.027	0.031	0.075
	N = 20,000	0.015	0.029	0.018	0.025	0.075	0.013	0.022	0.019	0.023	0.074
	N = 40,000	0.012	0.021	0.014	0.020	0.075	0.009	0.016	0.014	0.017	0.074

Note. RMSE = root mean squared error; BIC = Bayes information criterion.

were fairly accurate given the low RMSE values. The RMSE values were slightly lower for a longer test length of 21 items as compared with 14 items. However, differences were not very large. Generally, as the sample size increased, the RMSE value decreased. The RMSE values were substantially lower for 40,000 simulees as opposed to 5,000 simulees, especially for the dichotomous covariate gender and the dummy coded race/ethnicity covariate. Furthermore, not using the BIC stopping rule in the IRT-C procedure led to noticeably better estimates.

Estimated 3PLM Item Parameters

The accuracy of the estimated 3PLM item parameters were similar across medium and large DIF effect sizes. In view of this, we only present the medium DIF effect size results here (see Appendix for large DIF effect size results). Table 4 reveals that for both test lengths of 14 and 21, the estimates of the *a* and *b* item parameters were substantially better as compared with the pseudo-guessing *c* parameter. The correlation between estimated and true values for the *a* and *b* parameters were generally more than .90, but the correlation for the *c* parameter ranged from .58 to .87, corresponding to smaller and large sample sizes, respectively. As such, the *c* parameters are better estimated with a large sample size of 40,000. This is not surprising given that the *c* parameter is usually not as well estimated without specifying a prior distribution (Lord, 1986). A longer test length yielded marginally better estimates. There was little difference between using or not using the BIC stopping rule.

Type I Error and Power to Detect DIF

As shown in Table 5, for the medium DIF effect size, we found that the use of the BIC stopping rule led to overcontrolled Type I error rates. Specifically, the Type I error rates were substantially lower than the nominal rate of .05 and were closer to .005. On the other hand, without the BIC stopping rule, the Type I error rates were closer to the nominal rate of .05 at around .02 to .03. Importantly, across a range of sample sizes from 5,000 to 40,000, the Type I error rates did not fluctuate, showing that the IRT-C procedure produced very consistent control of Type I error rates.

The power to detect DIF was also substantially higher for the IRT-C procedure when the BIC stopping rule was not used. With the largest sample size of 40,000, with the BIC stopping criterion, the power to detect DIF was as low as .59 for both test lengths. Without the BIC stopping criterion, the power to detect DIF for a sample size of 40,000 was at least .80 for both test lengths. With this procedure there was moderate (.70) to high power (.98) to detect DIF across all the different covariates for sample sizes above 20,000. In general, there was also higher power to detect DIF with a longer test length.

Similar trends were found for the large DIF effect size condition as shown in Table 6. Type I error rates remained overcontrolled with the BIC stopping rule but close to the nominal rates without the BIC stopping rule. The power to detect DIF

Table 4. RMSE and Correlation Between Estimated and Simulated 3PLM Item Parameters for Medium DIF Effect Size.

		Nitem = 14				Nitem = 21			
		BIC stopping rule		No BIC stopping rule		BIC stopping rule		No BIC stopping rule	
Item parameter	Sample size	Correlation (*SD*)	RMSE (*SD*)	Correlation (*SD*)	RMSE (*SD*)	Correlation (*SD*)	RMSE (*SD*)	Correlation (*SD*)	RMSE (*SD*)
a	*N* = 5,000	0.863	0.335	0.862	0.319	0.905	0.212	0.902	0.252
		(0.094)	(0.647)	(0.108)	(0.549)	(0.053)	(0.083)	(0.040)	(0.401)
	N = 10,000	0.929	0.191	0.931	0.172	0.950	0.152	0.949	0.146
		(0.049)	(0.269)	(0.045)	(0.055)	(0.040)	(0.186)	(0.027)	(0.039)
	N = 20,000	0.964	0.121	0.961	0.121	0.974	0.100	0.974	0.100
		(0.026)	(0.04)	(0.026)	(0.034)	(0.013)	(0.022)	(0.013)	(0.021)
	N = 40,000	0.980	0.085	0.982	0.084	0.987	0.072	0.987	0.070
		(0.014)	(0.024)	(0.014)	(0.023)	(0.007)	(0.016)	(0.009)	(0.016)
b	*N* = 5,000	0.973	0.934	0.976	0.877	0.984	0.713	0.981	0.794
		(0.046)	(1.419)	(0.041)	(1.173)	(0.013)	(0.182)	(0.038)	(0.889)
	N = 10,000	0.988	0.699	0.989	0.657	0.990	0.653	0.991	0.638
		(0.016)	(0.537)	(0.007)	(0.128)	(0.016)	(0.367)	(0.005)	(0.098)
	N = 20,000	0.993	0.626	0.994	0.613	0.995	0.604	0.995	0.600
		(0.005)	(0.100)	(0.004)	(0.087)	(0.003)	(0.066)	(0.002)	(0.064)
	N = 40,000	0.995	0.596	0.996	0.583	0.996	0.589	0.996	0.583
		(0.003)	(0.068)	(0.002)	(0.062)	(0.002)	(0.050)	(0.002)	(0.047)
c	*N* = 5,000	0.542	0.077	0.523	0.077	0.580	0.071	0.588	0.070
		(0.243)	(0.023)	(0.227)	(0.023)	(0.197)	(0.018)	(0.184)	(0.018)
	N = 10,000	0.624	0.062	0.643	0.060	0.684	0.054	0.683	0.053
		(0.208)	(0.020)	(0.207)	(0.019)	(0.159)	(0.016)	(0.178)	(0.016)
	N = 20,000	0.764	0.043	0.742	0.045	0.781	0.041	0.788	0.039
		(0.158)	(0.016)	(0.192)	(0.017)	(0.129)	(0.012)	(0.131)	(0.013)
	N = 40,000	0.817	0.034	0.817	0.034	0.874	0.029	0.864	0.029
		(0.146)	(0.015)	(0.161)	(0.014)	(0.087)	(0.010)	(0.100)	(0.011)

Note. RMSE = root mean squared error; 3PLM = three-parameter logistic model; DIF = differential item functioning; BIC = Bayes information criterion.

Table 5. Medium DIF Effect Size: Power and Type I Error Rates for IRT-C Procedure for Test Lengths of 14 and 21 Items.

		Nitem = 14				Nitem = 21			
		BIC stopping rule		No BIC stopping rule		BIC stopping rule		No BIC stopping rule	
Sample size	Covariate	Power (*SD*)	Type I error (*SD*)	Power (*SD*)	Type I error (*SD*)	Power (*SD*)	Type I error (*SD*)	Power (*SD*)	Type I error (*SD*)
N = 5,000	Average	0.327	0.006	0.558	0.034	0.376	0.003	0.600	0.024
		(0.173)	(0.016)	(0.173)	(0.037)	(0.171)	(0.008)	(0.166)	(0.024)
	Gender	0.364	0.013	0.643	0.048	0.451	0.005	0.714	0.029
		(0.292)	(0.039)	(0.287)	(0.075)	(0.304)	(0.018)	(0.254)	(0.042)
	Race	0.028	0.001	0.240	0.018	0.037	0.000	0.284	0.014
		(0.104)	(0.009)	(0.225)	(0.045)	(0.100)	(0.004)	(0.241)	(0.029)
	Income	0.590	0.005	0.791	0.037	0.640	0.003	0.803	0.029
		(0.295)	(0.025)	(0.227)	(0.068)	(0.264)	(0.013)	(0.220)	(0.042)
N = 10,000	Average	0.474	0.005	0.748	0.034	0.528	0.002	0.757	0.023
		(0.196)	(0.016)	(0.153)	(0.036)	(0.199)	(0.006)	(0.141)	(0.022)
	Gender	0.591	0.011	0.848	0.04	0.686	0.003	0.874	0.028
		(0.332)	(0.044)	(0.203)	(0.066)	(0.300)	(0.013)	(0.175)	(0.041)
	Race	0.094	0.000	0.507	0.028	0.143	0.000	0.512	0.012
		(0.165)	(0.007)	(0.282)	(0.056)	(0.198)	(0.004)	(0.250)	(0.026)
	Income	0.736	0.003	0.889	0.035	0.754	0.002	0.884	0.028
		(0.269)	(0.016)	(0.16)	(0.056)	(0.263)	(0.011)	(0.161)	(0.041)

(continued)

Table 5. (continued)

		Nitem = 14				Nitem = 21			
		BIC stopping rule		No BIC stopping rule		BIC stopping rule		No BIC stopping rule	
Sample size	Covariate	Power (*SD*)	Type I error (*SD*)	Power (*SD*)	Type I error (*SD*)	Power (*SD*)	Type I error (*SD*)	Power (*SD*)	Type I error (*SD*)
N = 20,000	Average	0.667	0.002	0.855	0.031	0.705	0.001	0.863	0.027
		(0.203)	(0.011)	(0.119)	(0.037)	(0.186)	(0.004)	(0.112)	(0.023)
	Gender	0.814	0.003	0.936	0.036	0.839	0.002	0.943	0.033
		(0.261)	(0.020)	(0.129)	(0.062)	(0.238)	(0.010)	(0.119)	(0.045)
	Race	0.342	0.000	0.688	0.022	0.409	0.000	0.701	0.015
		(0.281)	(0.007)	(0.246)	(0.055)	(0.280)	(0.000)	(0.231)	(0.030)
	Income	0.846	0.003	0.941	0.035	0.868	0.001	0.946	0.033
		(0.214)	(0.019)	(0.115)	(0.062)	(0.195)	(0.007)	(0.120)	(0.043)
N = 40,000	Average	0.765	0.003	0.924	0.028	0.772	0.002	0.926	0.025
		(0.258)	(0.011)	(0.087)	(0.032)	(0.257)	(0.007)	(0.081)	(0.022)
	Gender	0.832	0.004	0.975	0.031	0.841	0.002	0.977	0.029
		(0.294)	(0.022)	(0.083)	(0.054)	(0.292)	(0.013)	(0.075)	(0.040)
	Race	0.591	0.000	0.814	0.019	0.594	0.000	0.822	0.012
		(0.347)	(0.005)	(0.208)	(0.044)	(0.344)	(0.000)	(0.188)	(0.028)
	Income	0.870	0.004	0.984	0.034	0.880	0.003	0.979	0.033
		(0.227)	(0.021)	(0.064)	(0.057)	(0.222)	(0.014)	(0.075)	(0.043)

Note. DIF = differential item functioning; IRT-C = item response theory with covariate; BIC = Bayes information criterion.

Table 6. Large DIF Effect Size: Power and Type I Error Rates for IRT-C Procedure for Test Lengths of 14 and 21 Items.

		Nitem = 14				Nitem = 21			
		BIC stopping rule		No BIC stopping rule		BIC stopping rule		No BIC stopping rule	
Sample size	Covariate	Power (*SD*)	Type I error (*SD*)	Power (*SD*)	Type I error (*SD*)	Power (*SD*)	Type I error (*SD*)	Power (*SD*)	Type I error (*SD*)
N = 5,000	Average	0.468	0.004	0.758	0.032	0.519	0.002	0.754	0.022
		(0.223)	(0.016)	(0.154)	(0.037)	(0.207)	(0.008)	(0.146)	(0.023)
	Gender	0.595	0.007	0.865	0.038	0.652	0.003	0.878	0.030
		(0.353)	(0.027)	(0.185)	(0.069)	(0.304)	(0.016)	(0.173)	(0.045)
	Race	0.138	0.001	0.544	0.023	0.178	0.000	0.525	0.011
		(0.205)	(0.013)	(0.274)	(0.048)	(0.220)	(0.004)	(0.256)	(0.025)
	Income	0.671	0.005	0.866	0.035	0.725	0.004	0.860	0.025
		(0.278)	(0.033)	(0.175)	(0.065)	(0.258)	(0.016)	(0.177)	(0.039)
N = 10,000	Average	0.650	0.002	0.861	0.029	0.672	0.002	0.860	0.024
		(0.247)	(0.008)	(0.125)	(0.033)	(0.242)	(0.009)	(0.121)	(0.021)
	Gender	0.760	0.003	0.934	0.035	0.776	0.004	0.944	0.030
		(0.310)	(0.018)	(0.137)	(0.060)	(0.311)	(0.016)	(0.116)	(0.041)
	Race	0.384	0.000	0.718	0.021	0.422	0.001	0.713	0.012
		(0.309)	(0.007)	(0.240)	(0.055)	(0.306)	(0.008)	(0.238)	(0.026)
	Income	0.807	0.002	0.929	0.031	0.817	0.002	0.924	0.029
		(0.247)	(0.013)	(0.128)	(0.055)	(0.235)	(0.012)	(0.142)	(0.041)

(continued)

Table 6. (continued)

		Nitem = 14				Nitem = 21			
		BIC stopping rule		No BIC stopping rule		BIC stopping rule		No BIC stopping rule	
Sample size	Covariate	Power (*SD*)	Type I error (*SD*)	Power (*SD*)	Type I error (*SD*)	Power (*SD*)	Type I error (*SD*)	Power (*SD*)	Type I error (*SD*)
N = 20,000	Average	0.787	0.002	0.917	0.030	0.828	0.001	0.920	0.025
		(0.243)	(0.012)	(0.090)	(0.036)	(0.211)	(0.005)	(0.088)	(0.024)
	Gender	0.857	0.003	0.968	0.037	0.904	0.001	0.971	0.030
		(0.274)	(0.020)	(0.097)	(0.070)	(0.215)	(0.009)	(0.083)	(0.042)
	Race	0.621	0.001	0.816	0.019	0.691	0.000	0.822	0.014
		(0.329)	(0.010)	(0.195)	(0.045)	(0.303)	(0.005)	(0.189)	(0.030)
	Income	0.882	0.003	0.966	0.033	0.889	0.001	0.967	0.031
		(0.216)	(0.022)	(0.094)	(0.061)	(0.205)	(0.012)	(0.088)	(0.041)
N = 40,000	Average	0.749	0.002	0.954	0.030	0.748	0.003	0.951	0.025
		(0.322)	(0.011)	(0.069)	(0.032)	(0.327)	(0.012)	(0.064)	(0.022)
	Gender	0.786	0.003	0.989	0.035	0.787	0.003	0.990	0.030
		(0.350)	(0.016)	(0.059)	(0.057)	(0.362)	(0.015)	(0.049)	(0.041)
	Race	0.631	0.001	0.884	0.018	0.626	0.001	0.875	0.012
		(0.395)	(0.010)	(0.164)	(0.043)	(0.391)	(0.010)	(0.160)	(0.027)
	Income	0.831	0.003	0.987	0.037	0.832	0.004	0.987	0.031
		(0.276)	(0.018)	(0.059)	(0.061)	(0.276)	(0.021)	(0.056)	(0.041)

Note. DIF = differential item functioning; IRT-C = item response theory with covariate; BIC = Bayes information criterion.

was higher as expected with a larger DIF effect size. Without the BIC stopping criterion, the power to detect DIF for a sample size of 20,000 was at least .80 for both test lengths. There was at least moderate power (.70) to detect DIF across all the covariates for sample sizes above 10,000.

Discussion

In typical large-scale testing contexts, it is important to examine whether DIF occurs across demographic groups for sociopolitical and legal reasons. Nevertheless, demographic categories are often imperfect proxies for underlying psychological characteristics that may lead to DIF. For example, DIF found between African Americans and Caucasians may be the result of multiple causes including socioeconomic status, cultural experiences, and language use. For scientific purposes, understanding the proximal reasons for DIF is important. The IRT-C procedure for multiple covariates can help identify sociological and psychological reasons underlying demographic differences.

In the first simulation study to assess the IRT-C procedure for detecting DIF simultaneously across multiple covariates, initial promising results were found for the IRT-C procedure without the BIC stopping rule. With a reasonably large sample size of about 20,000, there was good covariate and item parameter recovery, although a sample size of 40,000 was needed for good item parameter recovery of the *c* parameter. Furthermore, Type I error rates were consistently well-controlled across different sample sizes and reasonably close to the nominal Type I rate of .05. There was also moderate to good power to detect DIF for a sample size of 20,000 or more across the different covariates with the moderate DIF effect size condition; a small sample size of 10,000 was needed for the large DIF effect size condition. This demonstrates that the IRT-C procedure can likely be used in large scale testing contexts to examine DIF across multiple variables.

While this was not a focus of our study, another advantage of the IRT-C model is that we can examine the standardized latent mean difference between demographic groups where DIF is modeled versus when it is not modeled. This can be helpful in determining the extent DIF can potentially affect latent mean differences. More research can examine whether this can be used to evaluate DIF effect sizes in a heuristic fashion where latent means with and without DIF being estimated are compared (see Tay, Meade, & Cao, 2015).

Limitations

There are several limitations to the current study and the IRT-C procedure as currently implemented in our study. The IRT-C procedure is currently limited to more than 5,000 respondents. While this sample size requirement may not be a substantial problem in large-scale testing, it may pose a problem when seeking to implement this in small-scale testing. Perhaps a bigger problem in implementing the IRT-C procedure is the possible limitation on test length in which computational requirements place constraints on how quickly one can estimate longer test lengths (e.g., 35 items). While this may not be prohibitive for any single estimation in an applied setting, we were practically unable to assess the IRT-C procedure for test lengths substantially longer than 21 items. However, extrapolating from our current study, it is likely that parameter recovery would be improved along with higher power to detect DIF with a longer test length.

Future Research and Conclusion

The current study focuses on the detection of uniform DIF across multiple covariates and there are several areas for future research. First, research can examine the extent nonuniform DIF can be detected using the IRT-C procedure. Initial examination in the context of a two-group comparison showed that while the Type I error rates are controlled close to the nominal rate, the power to detect nonuniform DIF is lower than uniform DIF (Tay et al., 2013). Second, in this study we only simulated the least complex case when there are main covariate effects leading to DIF rather than interactions between covariates leading to DIF. More research needs to determine whether covariate interaction-related DIF can be detected with sufficient power. Third, more research can also examine other indices for flagging DIF apart from the BVR as it may affect power to detect DIF (Tay et al., 2013). Fourth, research needs to be conducted on estimating an IRT-C procedure alongside the use of *c* prior distributions to determine whether estimates and DIF detection can be improved. Finally, the iterative IRT-C procedure took a substantial time to finish—one replication for a sample size of 40,000 took around 22 to 30 hours. More research can assess whether a reduced iterative procedure of evaluating DIF on multiple items at each iteration can accomplish similar results.

In conclusion, the simulation results for the IRT-C procedure points to the unique potential of identifying DIF on the 3PL models across multiple covariates, even with a large proportion of DIF items as high as 50%. Without the BIC stopping criterion, Type I error rates were well controlled for using the Wald test where the correct asymptotic sampling variance–covariance matrix is used. This study has demonstrated the viability of the IRT-C procedure for assessing DIF in large scale testing where there are large sample sizes and long test lengths. We hope that this set of promising results can engender the use of the IRT-C model and procedure, and also increased research in this burgeoning area.

Appendix

Table A.1. RMSE of Estimated and Simulated Covariate Coefficients for Large DIF Effect Size.

		BIC stopping rule (RMSE)					No BIC stopping rule (RMSE)				
Test length	Sample size	Female	Black	Hispanic	Asian	ZIncome	Female	Black	Hispanic	Asian	ZIncome
Nitem = 14	*N* = 5,000	0.048	0.082	0.039	0.073	0.083	0.035	0.057	0.043	0.055	0.076
	N = 10,000	0.030	0.056	0.028	0.046	0.078	0.023	0.035	0.027	0.038	0.075
	N = 20,000	0.022	0.036	0.020	0.034	0.077	0.016	0.024	0.020	0.027	0.075
	N = 40,000	0.026	0.037	0.014	0.031	0.080	0.011	0.018	0.014	0.018	0.074
Nitem = 21	*N* = 5,000	0.035	0.059	0.037	0.057	0.080	0.027	0.049	0.036	0.048	0.076
	N = 10,000	0.026	0.042	0.027	0.037	0.077	0.019	0.029	0.026	0.033	0.075
	N = 20,000	0.015	0.028	0.020	0.026	0.075	0.014	0.022	0.019	0.025	0.074
	N = 40,000	0.020	0.027	0.014	0.024	0.078	0.010	0.015	0.013	0.016	0.074

Note. RMSE = root mean squared error; DIF = differential item functioning; BIC = Bayes information criterion.

Table A.2. RMSE and Correlation Between Estimated and Simulated 3PLM Item Parameters for Large DIF Effect Size.

		Nitem = 14				Nitem = 21			
		BIC stopping rule		No BIC stopping rule		BIC stopping rule		No BIC stopping rule	
Item parameter	Sample size	Correlation (*SD*)	RMSE (*SD*)	Correlation (*SD*)	RMSE (*SD*)	Correlation (*SD*)	RMSE (*SD*)	Correlation (*SD*)	RMSE (*SD*)
a	*N* = 5,000	0.849	0.412	0.856	0.359	0.898	0.246	0.897	0.276
		(0.129)	(0.887)	(0.107)	(0.732)	(0.070)	(0.383)	(0.088)	(0.521)
	N = 10,000	0.922	0.197	0.929	0.184	0.949	0.147	0.950	0.145
		(0.055)	(0.26)	(0.052)	(0.234)	(0.026)	(0.037)	(0.026)	(0.037)
	N = 20,000	0.963	0.122	0.961	0.123	0.974	0.103	0.975	0.100
		(0.026)	(0.033)	(0.031)	(0.039)	(0.014)	(0.024)	(0.013)	(0.024)
	N = 40,000	0.976	0.094	0.981	0.086	0.984	0.077	0.987	0.071
		(0.018)	(0.031)	(0.014)	(0.027)	(0.010)	(0.02)	(0.007)	(0.017)
b	*N* = 5,000	0.967	1.104	0.975	0.976	0.980	0.803	0.978	0.843
		(0.062)	(1.939)	(0.046)	(1.569)	(0.040)	(0.938)	(0.052)	(1.111)
	N = 10,000	0.986	0.723	0.988	0.694	0.990	0.650	0.991	0.636
		(0.015)	(0.638)	(0.021)	(0.623)	(0.006)	(0.102)	(0.006)	(0.101)
	N = 20,000	0.992	0.633	0.994	0.624	0.994	0.607	0.995	0.599
		(0.007)	(0.095)	(0.004)	(0.099)	(0.004)	(0.070)	(0.003)	(0.062)
	N = 40,000	0.992	0.614	0.995	0.593	0.994	0.598	0.996	0.585
		(0.008)	(0.076)	(0.003)	(0.060)	(0.004)	(0.059)	(0.002)	(0.050)
c	*N* = 5,000	0.546	0.075	0.553	0.075	0.568	0.070	0.579	0.069
		(0.241)	(0.023)	(0.241)	(0.023)	(0.203)	(0.019)	(0.190)	(0.018)
	N = 10,000	0.650	0.058	0.661	0.058	0.696	0.053	0.693	0.052
		(0.213)	(0.021)	(0.207)	(0.020)	(0.165)	(0.016)	(0.162)	(0.015)
	N = 20,000	0.728	0.045	0.739	0.045	0.787	0.040	0.786	0.039
		(0.188)	(0.018)	(0.190)	(0.018)	(0.135)	(0.013)	(0.135)	(0.013)
	N = 40,000	0.807	0.037	0.831	0.033	0.863	0.029	0.866	0.029
		(0.152)	(0.017)	(0.132)	(0.014)	(0.104)	(0.011)	(0.098)	(0.011)

Note. RMSE = root mean squared error; 3PLM = three-parameter logistic model; DIF = differential item functioning; BIC = Bayes information criterion.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. We used SAT data to mimic the demographic characteristics in generating score distributions. We used simulated ability estimates, not observed SAT examinee scores. The item and DIF parameters used in the simulation were also simulated as will be described and do not reflect the SAT in any way. SAT® is a registered trademark of the College Board, which was not involved in the research or development of the manuscript.

References

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*, *36*, 185-198.

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

Langer, M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation* (Unpublished doctoral dissertation). University of North Carolina, Chapel Hill.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, *23*, 157-162.

Robert, C., Lee, W. C., & Chan, K.-Y. (2006). An empirical analysis of measurement equivalence with the INDCOL measure of individualism and collectivism: Implications for valid cross-cultural inference. *Personnel Psychology*, *59*, 65-99.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*, 1292-1306. doi:10.1037/0021-9010.91.6.1292

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*, 3-46.

Tay, L., Newman, D. A., & Vermunt, J. K. (2011). Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence. *Organizational Research Methods*, *14*, 147-176. doi:10.1177/1094428110366037

Tay, L., Vermunt, J. K., & Wang, C. (2013). Assessing the item response theory with covariate (IRT-C) procedure for ascertaining DIF. *International Journal of Testing*, *13*, 201-222. doi: 10.1080/15305058.2012.692415

Vermunt, J. K., & Magidson, J. (2008). Latent GOLD 4.5 [computer program]. Belmont, MA: Statistical Innovations.

Woods, C. M. (2009a). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*, 42-57.

Woods, C. M. (2009b). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*, 1-27.

Woods, C. M., Cai, L., & Wang, M. (2012). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*. Advance online publication. doi:10.1177/0013164412464875

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

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2009). Illustration of MIMIC-model DIF testing with the Schedule for Nonadaptive and Adaptive Personality. *Journal of Psychopathology and Behavioral Assessment*, *31*, 320-330.