Running head: ASSESSING DIF WITH IRT-C

Assessing the item response theory with covariate (IRT-C) procedure for ascertaining differential item functioning

Louis Tay

University of Illinois at Urbana-Champaign

Jeroen K. Vermunt

Tilburg University

Chun Wang

University of Illinois at Urbana-Champaign

Address for correspondence:
Louis Tay
Department of Psychology
University of Illinois at Urbana-Champaign
603 East Daniel Street
Champaign, Illinois 61820
Email: sientay@illinois.edu

Abstract

We evaluate the item response theory with covariates (IRT-C) procedure for assessing DIF without preknowledge of anchor items (Tay, Newman, & Vermunt, 2011). This procedure begins with a fully constrained baseline model and candidate items are tested for uniform and/or non-uniform DIF using the Wald statistic. Candidate items are selected in turn based on high unconditional bivariate residual (UBVR) values. This iterative process continues until no further DIF is detected or the Bayes information criterion (BIC) increases. We expanded on the procedure and examined the use of conditional bivariate residuals (CBVR) to flag for DIF; aside from the BIC, alternative stopping criteria were also considered. Simulation results showed that the IRT-C approach for assessing DIF performed well, with the use of CBVR yielding slightly better power and Type I error rates than UBVR. Additionally, using no information criterion yielded higher power than using the BIC although Type I error rates were generally well controlled in both cases. Across the simulation conditions, the IRT-C procedure produced results similar to the Mantel-Haenszel and MIMIC procedures.

Assessing the item response theory with covariate (IRT-C) procedure for ascertaining differential item functioning

Differential item functioning (DIF) occurs when the expected item score conditioned on the latent trait differs due to group membership. Because this issue is important to the fairness of psychological and educational tests, or the equivalence of scores in cross-cultural settings, a variety of methods have been proposed for assessing DIF (see review by Millsap & Everson, 1993). These include the Mantel-Haenszel approach (Holland & Thayer, 1988), the logistic regression method (Swaminathan & Rogers, 1990), the chi-square method (Lord, 1980), differences in area between item response curves (Raju, 1988), and likelihood-ratio test methods (Thissen, Steinberg, & Wainer, 1993).

Although these methods for assessing DIF have conventionally been used, they can be limited when seeking to understand the sources of DIF in an international setting. Often, DIF is tested between countries when using these methods -- this implicitly assumes that country membership is the source of DIF. However, country membership may be a proxy for other differences such as wealth, rural-urban settings, language, or simply differences in the demography of a sample (Matsumoto & Yoo, 2006). In order to accurately assess the sources of DIF, or to control for potential confounds, we need to use methods that can incorporate multiple observed characteristics simultaneously. For this purpose, two methods have been proposed: the multiple indicators multiple cause (MIMIC) model for assessing DIF in a confirmatory factor analytic framework (Muthén, 1985, 1988) and the item response theory with covariates (IRT-C) model (Tay, Newman, & Vermunt, 2011).

For detecting DIF, the utility of both MIMIC and IRT-C models are: (a) one can examine DIF across multi-categorical and continuous covariates; (b) DIF on multiple covariates can be examined simultaneously; (c) differences in latent means and variances across groups can be estimated and compared in the final model. Therefore, unlike conventional measurement

equivalence procedures utilizing multiple group comparisons (e.g., multiple group confirmatory factor analysis or IRT), potential sources of DIF and mean level differences can be examined in concert. For example, we can ascertain the degree to which DIF is attributable to gender, race, language, and/or socioeconomic status; and after controlling for sources of nonequivalence, the degree to which mean level differences are linked to gender, race, language, and/or socioeconomic status. Consequently, important sources of DIF can be identified and controlled simultaneously.

In MIMIC models, an IRT model is fit to a matrix of polychoric correlations thereby using limited information. In IRT-C models, an IRT model is directly fit to the data thus utilizing full information. There are no large differences between limited information and full information methods in the estimation of parameters and standard errors (Forero & Maydeu-Olivares, 2009). However, full information methods allow for more complex models (e.g., mixture models) which at times may not be identified with limited information methods (Bolt, 2005). For example, a recent DIF study specified a three-parameter logistic model using the IRT-C approach (Tay, Drasgow, & Vermunt, 2011).

Simulation studies have examined the power and Type I error rates of MIMIC models for assessing uniform and non-uniform DIF (e.g., Finch, 2005; Shih & Wang, 2009; Woods, 2009a; Woods & Grimm, in press), but the proposal of an IRT-C model for assessing DIF (Tay, Newman, & Vermunt, 2011) is recent and there has not been a systematic examination of such an approach. Some simulations with MIMIC models have generally assumed that anchor items are known *a priori* (Woods, 2009a; Woods & Grimm, in press). In such a procedure, a MIMIC model is specified such that anchor items are constrained to be equal across groups while DIF parameters are estimated and tested with the Wald statistic (Woods, 2009a), or the likelihood-ratio statistic where the constrained model is tested against the baseline (Thissen et al., 1993).

An IRT-C model can also be specified using such a procedure (Tay, Drasgow, & Vermunt, 2011) producing well-controlled Type I error rates and high power for detecting DIF.

However, an issue commonly encountered in the detection of DIF is that anchor items between groups are not known before hand. Although there are ways to identify anchor items in two-group comparisons in MIMIC models (Woods, 2009b), identifying anchors across multiple covariates (e.g., race, gender, and socioeconomic status) is complicated. Indeed, we have argued that the primary utility of the MIMIC and IRT-C model is to identify DIF across multiple covariates simultaneously. In view of this, researchers have been examining the use of approaches that do not require known anchor items to MIMIC models but instead identify anchors through a scale purification approach (Wang, Shih & Yang, 2009). Following this line of research, the proposed IRT-C model and procedure (Tay, Newman & Vermunt, 2011) is used for assessing DIF without preknowledge of anchor items and it can be generalized to instances when there are multiple covariates.

In this study, we examine the power and Type I error rates of the IRT-C approach when anchor items are unknown for assessing DIF. We compare this procedure to the Mantel-Haenszel approach and the MIMIC approach where anchor items are assumed to be known (Woods, 2009a; Woods & Grimm, in press). We focus only on the simplest case: a two group comparison commonly undertaken in DIF studies. Indeed, the robustness of this model in a two group instance is a minimal hurdle for conducting studies on more generalized cases (e.g., multiple groups, or assessing multiple covariates simultaneously).

The IRT-C model

A 2-parameter logistic model (2PLM) is utilized to describe the relationship between the probability of item endorsement and the latent trait level θ_j. Let y_{ji} denote the response of individual j, $j = 1,\ldots,J,$ on item i, $i = 1,\ldots,I$; the probability of item endorsement is then

$$P(y_{ji} \mid \theta_j) = \frac{1}{1+\exp(-[a_i\theta_j + b_i])}, \quad (1)$$

where a_i and b_i represent the item discrimination and item location respectively. DIF occurs when the expected score given the same latent trait θ_j is different by virtue of observed characteristic (z_j) (Hulin, Drasgow, & Parsons, 1983). For instance, an observed characteristic may be gender or race. DIF can be represented as

$$P(y_{ji} \mid \theta_j, z_j) = \frac{1}{1+\exp(-[a_i\theta_j + b_i + c_i z_j + d_i z_j \theta_j])}, \quad (2)$$

where the probability of item responding depends not only on θ_j but also on z_j. The additional terms in equation (2), $c_i z_j$ and $d_i z_j \theta_j$ represent the direct and interaction effects for modeling uniform and non-uniform DIF, respectively. The significance of the terms c_i and d_i from a likelihood-ratio test can be used to ascertain whether uniform and/or non-uniform DIF occurs. Because freely estimating the c_i and d_i terms for every item across an observed characteristic (z_j) would lead to an unidentified model, the DIF procedure proposed is such that in the initial model, all the c_i and d_i terms are constrained to zero, leading to a fully constrained initial model. These coefficients are only estimated and tested for significance when DIF is likely to be found on an item. We elaborate on the IRT-C procedure for detecting DIF in the next section.

One can extend equation (2) by including a vector of observed characteristics $\underset{\sim}{z}_j$, or covariates, to include multiple observed categories and/or continuous variables. Thus, we can determine if DIF occurs across multiple groups (e.g., Kim, Cohen, & Park, 1995; Penfield, 2001) or across (continuous) covariate patterns. In this paper, however, we focus only on the two-group case (i.e., a referent and a focal group) to determine whether the IRT-C model can be

used to detect DIF in a traditional manner because it is not known how well such a procedure fares in the first place.

In the IRT-C model, the distributions of the latent traits across the observed groups are modeled as well in the testing of DIF as given by

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

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

At this juncture, we seek to clarify differences with some DIF procedures. First, we note that the IRT-C model is not merely an algebraic manipulation, but stems from a general latent variable modeling framework which enables the estimation of the additional coefficients associated with DIF (i.e., c_i and d_i) unachievable in prior IRT models. Second, the likelihood ratio test for DIF in the IRT-C model pertains to the coefficients associated with DIF (i.e., c_i and d_i) and is different from the likelihood ratio test (Thissen, Steinberg, & Wainer, 1993) between two models: the unconstrained and constrained models. Third, the IRT-C model is different from the logistic regression approach for testing DIF. In the IRT-C model, the observed score on an item varies as a function of the latent trait θ as shown in Equation 2. For logistic regression, the observed score on an item varies as a function of the total test score *x*.

An examination of three DIF Procedures

The IRT-C procedure. As mentioned earlier, freely estimating all the coefficients associated with DIF (i.e., c_i and d_i) would lead to an unidentified model. Therefore, the IRT-C procedure for assessing DIF begins with a fully constrained model in which all these coefficients are constrained to zero. Instead, the IRT-C procedure utilizes unconditional

bivariate residuals (UBVR) or conditional bivariate residuals (CBVR) to identify candidate DIF items. The equations for the UBVR and CBVR are shown in the Appendix. For values with large UBVRs, the parameters associated with DIF (i.e., c_i or d_i in Equation 2) for a specific covariate are then estimated and tested for significance using the Wald statistic.

A general description of the IRT-C procedure is as follows. A fully constrained baseline model is first specified in which all the items are set as invariant between the groups, or more generally, the (continuous) covariate patterns. Then, an iterative procedure is used to identify the presence of uniform and non-uniform DIF. At each stage, only the item with the largest UBVR or CBVR is tested for DIF using the Wald statistic. If DIF occurs, the initial fully constrained model is updated by allowing the identified item to be freely estimated across both groups. Using the updated model, the item flagged by UBVR or CBVR is tested for DIF. This process continues until no statistically significant DIF occurs or the model is less parsimonious as indicated from an information criterion.

Starting from a fully constrained baseline model, the algorithm implemented is as follows:

(a) For the current model (M0), the possible presence of DIF is determined by examining the UBVR or CBVR between each covariate and each indicator. The largest BVR is flagged and tested for DIF. For instance, if there are two covariates and 10 indicators, the item with the largest BVR in the 2 × 10 BVR matrix is flagged as possibly having DIF. See Appendix for an illustration.

(b) In a subsequent model (M-NU [non-uniform]), both uniform and non-uniform DIF for the flagged item are specified; that is c_i and d_i parameters are freely estimated across groups. The statistical significance for non-uniform DIF d_i is determined using the Wald statistic. If the non-uniform DIF is significant, both c_i and d_i will continue to be freely estimated and step (c) is skipped. Otherwise, only uniform DIF is examined with a subsequent model in step (c).

(c) In this model (M-U [uniform]), only uniform DIF is specified for the flagged item and examined for statistical significance using the Wald statistic. If the uniform DIF parameter c_i is not significant, no DIF is found on the flagged item; the iterative procedure ends and M0 is selected as the final model. Otherwise, the item parameter c_i will continue to be freely estimated.

(d) If the interim models M-NU or M-U have BIC values that are higher than M0, the iterative procedure ends, and M0 is selected as the final model. To elaborate, if the BIC values are higher for the interim models, the original M0 model is preferred even if the c_i and d_i parameters are significant. This is because the model with additional DIF is less parsimonious than M0. Otherwise, if the interim models have lower BIC values than M0, step (e) is undertaken.

(e) The IRT-C is reestimated keeping only significant DIF effects from steps (b) and/or (c); that is, either M-NU or M-U replaces M0. Hence, the original model is updated with the interim model M-NU or M-U. This new model is then used to test for DIF in subsequent items. Steps (a) through (d) are then repeated until no DIF occurs as evaluated in step (c) or if the BIC values rises higher than the penultimate model (M0) as evaluated in step (d).

The above procedure describes the stepwise process outlined by Tay et.al. (2011). In the initial proposal, only UBVRs were used to illustrate the testing of DIF. In our simulations, we examine several variants of this IRT-C procedure. First, we investigate the use of CBVRs and compare it to the UBVR. Second, we compare two different stopping criteria: (i) we substitute BIC with AIC3. Past research has shown that the AIC3 performs well for model selection with multivariate categorical responses: it is slightly more liberal than AIC and more conservative than BIC (Dias, 2004; Vermunt, van Ginkel, van der Ark, & Sijtsma, 2008). (ii) Instead of using the BIC, we do not use any information criteria and the iterative process stops only when no DIF occurs on the item (tested by the Wald statistic) with the largest BVR value. Therefore,

our simulations test three different stopping rules: no information criterion, AIC3, and BIC. In general, it is expected that the no information criterion condition would have higher power to detect DIF but also higher Type I error rates because the iterative procedure continues until no more DIF is detected; however, in the AIC3 and BIC condition, the iterative procedure ends a less parsimonious model is utilized even though DIF was detected.

This proposed procedure is essentially a type of scale purification process. Wang, Shih, and Yang (2009) used a scale purification procedure with the MIMIC model. The analytic steps are as follows: First, using a fully constrained baseline model, they proposed that each item be tested for DIF in turn. Second, non-DIF items are used as anchor items. Third, all non-anchor items are tested for DIF. If all non-anchor items have DIF, the process ends. Otherwise, the second step (in this case, the set of anchor items is updated by including non-anchor items that do not exhibit DIF) and third step is repeated until all non-anchor items can be shown to have DIF. The procedure by Tay, Newman and Vermunt (2011) is different in that DIF is only tested for candidate items that have a large BVR. This limits the number of items that are tested for DIF.

Importantly, because we are concerned with the practical testing of DIF, the procedure encompasses a type of scale purification, or the identification of anchor items. Therefore, our study merges two issues in the IRT-C procedure: a scale purification approach to identify anchor items (i.e., invariant items), and DIF detection with the IRT-C model when anchor items are identified. Our preference is to examine this in a single study as compared to two separate studies. This is because in practice, the set of anchor items that are invariant across groups is usually unknown. If a set of anchor items is known, it would be possible to estimate the coefficients associated with DIF (i.e., c_i and d_i) for all the non-anchor items and determine their significance in a single step (e.g., Tay, Drasgow, & Vermunt, 2011).

Mantel-Haenszel procedure. For the purpose of comparison, we calculated the Mantel-Haenszel (MH) statistic to determine its power and Type I error rate for detecting DIF. The MH statistic was based on the formula from Narayanan and Swaminathan (1994). No correction for discontinuity was used because previous research has shown that this produces better control of Type I error rates (Paek, 2010). The MH statistic is considered because it is the most widely used and simplest DIF detection technique. More importantly, this procedure does not rely on any specific item response model. By only utilizing 2 by 2 contingency table, MH provides a useful benchmark for the current investigation. We are interested to see whether the IRT-C method, by using fully parametric approach, will perform better.

MIMIC procedure. We used the MIMIC procedure described by Woods (2009a). In this procedure, it is presumed that a set of items decided from preliminary tests are designated as invariant across groups; these invariant items are termed "designated anchors". A free-baseline model is specified such that all items are presumed to have DIF (with the exception of the designated anchors). The log-likelihood of the free-baseline model is subsequently compared to the log-likelihood of more constrained models where each studied item, formerly presumed to have DIF, is constrained to be equivalent across groups. In this manner, DIF for each item can be evaluated using the -2 difference in the log-likelihoods between the free-baseline and constrained model, which has an approximate χ^2-distribution with 1 degree-of-freedom.

For the purposes of our study, we prefer the designated anchor MIMIC approach as compared to the scale purification MIMIC approach. This is because we want to compare the IRT-C procedure to the best possible MIMIC alternative. Possible inaccuracies from scale purification may lower the accuracy of the MIMIC procedure.

Summary

Tay, Newman, and Vermunt (2011) illustrated the use of IRT-C to assess DIF: items that have high UBVRs are tested for DIF iteratively until a stopping criteria based on the BIC is

reached. We examine several variants of the IRT-C procedure (2 BVR types × 3 stopping criteria). We use a simulation study to evaluate the power and Type I error rates for assessing DIF with the IRT-C approach and compared it to the Mantel-Haenszel and MIMIC procedures. In addition, an advantage of the IRT-C approach is the ability to test for specific types of DIF (uniform or non-uniform). We evaluate whether the IRT-C approach has sufficient power to detect specific types of DIF. Also, because the IRT-C approach estimates the focal group latent mean, we examine how well this is recovered with the root-mean-squared-error (RMSE).

Method

Simulation parameters. In an initial study, we examined conditions in which we had test lengths of 10 and 20 items, and sample sizes of 250, 500, 1000, and 2000. Because the results were consistent, we used a fixed sample size of 500 respondents in the reference group and a scale length of 10 items in this study. This sample size and test length used was consistent with conditions used in past simulation studies examining traditional DIF methods (e.g., Meade et al., 2007; Stark et al., 2006) and the MIMIC model (e.g., Woods, 2009a). We chose to focus on other factors that may affect DIF detection; these included the examination of different focal group sample sizes, percentage of DIF, the various types of DIF, and latent mean differences between the reference and focal group. Readers interested in finding out more about the initial findings can require for more details from the corresponding author.

The simulation parameters varied in this study were focal group sample size (250 or 500), percentage of DIF (20% or 40%), DIF conditions (11 conditions which will be explicated). Altogether, there were a total of 44 simulation conditions. Initial simulations where we included (non-)differences in latent means between groups ($\theta_{mean.focal}$= 0, or $\theta_{mean.focal}$=-0.3) and (non-)difference in latent trait standard deviations ($\theta_{sd.focal}$ = 1, $\theta_{sd.focal}$ = .8) were shown to be similar. Hence, we did not vary differences in the latent standard deviations between the reference and focal group and the latent mean for the focal group was set to -.30. Within each

condition, 200 simulations were undertaken for each of procedure: the IRT-C (all 6 variations), the Mantel-Haenszel, and the MIMIC procedure.

Data generation. For a traditional 2PLM, given by

$$P(y_{ji} \mid \theta_j) = \frac{1}{1+\exp(-1.702a_i^*[\theta_j - b_i^*])}, \quad (4)$$

item discriminations a_i^* are 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; and b_i^* values are sampled from a uniform (-2, 2) distribution. Final generated item parameters a_i and b_i are given by transformations

$$a_i = 1.702 \times a_i^* \text{ and } b_i = -1.702 \times a_i^* \times b_i^*$$

These values were taken to emulate item parameters commonly encountered in self-reported typical behaviors.

Differential item functioning conditions. Apart from a condition where we did not simulate DIF, we simulated 10 different other DIF conditions which varied with respect to three factors (a) the occurrence of uniform and/or non-uniform DIF, (b) the amount of DIF (small or large), and (c) whether DIF was non-compensatory or compensatory. Table 1 presents how the three design factors were crossed to produce the 10 DIF conditions.

Non-compensatory DIF: Following similar procedures in emulating DIF used in previous 2PLM (see equation 12) simulations, non-uniform DIF was simulated by deducting a value of .40 from a_i^* for the focal DIF items (e.g., Finch & French, 2008). Two types of uniform DIF were simulated: small or large uniform DIF effects were specified by adding a value of .40 or .80, respectively, to the b_i^* focal DIF items (e.g., Rogers & Swaminathan, 1993).

Compensatory DIF: Similar to procedures used by Meade and colleagues (2007), compensatory DIF was simulated by specifying DIF that occurs in opposing directions for half

of the DIF items. For a scale length of 10 in which both uniform and small non-uniform DIF are specified, a value of .40 would be subtracted from a_i^* for one DIF item, but added to a_i^* for another DIF item; similarly, a value of .40 would be subtracted from b_i^* for one DIF item, but added to the b_i^* value for another DIF item.

In both the IRT-C and Mantel-Haenszel procedures, DIF items were randomly selected – for a test length of 10 items – from all items; by contrast, because the first item on the scale was the designated anchor item for the MIMIC procedure, the choice of DIF items fell on the remaining 9 items. We note that this led to slightly different datasets between the former procedures and the MIMIC procedure. However, we did not think that this would lead to systematic differences because all the items were randomly generated in each replication. After simulating DIF on the selected items using the traditional 2PLM, these DIF item parameters were then transformed into the 2PLM metric. The software Latent GOLD was used to generate the data based on the specified item parameters and latent distributions for the comparison groups piped from the statistical software R (2008).

Estimation. Latent GOLD was used for the maximum likelihood (ML) estimation of item parameters and latent trait means and standard deviations, as well as for the computation of the UBVR and CBVR values. With the exception of increasing the number of quadrature points from the standard of 10 to 50, all the default settings were used.

Results

The simulation results showed that the IRT-C procedure, using the AIC3 stopping criterion produced results (i.e., power, Type I error, and RMSE of the estimated focal latent mean) that were very similar to those of the no information criterion condition. This implied that using the AIC3 to select a final model coincided with that of the no information criterion rule. Further, the UBVR and the CBVR produced similar results although the CBVR had

slightly higher power; across the simulation conditions, the power for CBVR is on average higher by around .04 and there is no difference in the averaged Type I error rates. Thus, we only present illustrative results where the CBVR was used across two stopping criteria: no information criteria and the BIC. Tables 1 to 4 present the simulation results in which we compare the power and Type I error rates of the IRT-C procedure with the Mantel-Haenszel and MIMIC procedures.

Comparing IRT-C stopping rules. When there was a moderate proportion of DIF items on the test (i.e., percentage of DIF was 20%), the no information criterion stopping rule produced Type I error rates that were closest to the nominal Type I error rate of .05 whereas the BIC stopping criterion had Type I error rates close to .01. By contrast, when there was a large proportion of DIF items (i.e., percentage of DIF was 40%), the BIC stopping criterion had Type I error rates close to .05 and the no information criterion stopping rule had a higher Type I error rate that averaged .10. In both conditions however, the power to detect DIF was higher for the no information criterion stopping rule. This suggests that the IRT-C procedure with no information criterion performed relatively better.

Comparing IRT procedures. When there was a moderate proportion of DIF items on the test, the IRT-C procedure with no information criterion performed the best in that the Type I error rates were well controlled and there was high power relative to the other procedures. The Mantel-Haenzsel procedure had elevated Type I error rates around .10 and similar power to the IRT-C procedure with no information criterion. The MIMIC model had Type I error rates that were close to the nominal Type I error rates but power was slightly lower than the IRT-C procedure. In particular, when only non-uniform DIF was present (conditions 2 and 3 in Tables 1 to 2), the IRT-C procedure without the use of information criteria substantially outperformed the MIMIC procedure for detecting DIF. For example, Table 1 shows that the power for the IRT-C procedure was around 0.55 whereas power for the MIMIC procedure was around 0.25.

When there were a large proportion of DIF items on the scale, the MIMIC procedure outperformed the IRT-C procedure with no information criterion. For the IRT-C procedure, the Type I error rates were on average close to .10 but the Type I error rates for the MIMIC procedure were well controlled at 0.05. On average, these two procedures yielded similar power across the various DIF conditions. Interestingly, when DIF was compensatory, the IRT-C procedure with no information criterion generally had higher power than the MIMIC procedure and Type I error rates close to the nominal Type I error rate as well. This demonstrated that the IRT-C procedure was better when DIF was compensatory. Both the MIMIC and IRT-C procedures outperformed that Mantel-Haenzsel procedure because of its high Type I error rates.

Power to detect specific forms of DIF. In addition to detecting DIF on an item, we evaluated the power of the IRT-C procedure to detect specific forms of DIF (i.e., uniform or non-uniform) as shown in the column "Power.Form". Overall, the power to detect the exact form of DIF on an item was lower than the power to detect any type of DIF. Nevertheless, this was because the power to specifically detect non-uniform DIF was substantially lower; power to specifically detect uniform DIF was comparable to power for detecting any type of DIF.

RMSE of the estimated focal latent mean. Overall, the IRT-C procedure as implemented in Latent GOLD 4.5 accurately estimates differences in the latent means when DIF is accurately accounted for. When there was a moderate proportion of DIF on the test, the RMSEs of the focal latent mean for the IRT-C procedure were low and consistent across all the simulated conditions at around .10. When there was a large proportion of DIF on the test, the average RMSE was 0.15 for the IRT-C procedure as compared to the former 0.10. This is expected because accurately detecting DIF in items (when there is a moderate proportion of DIF items) in turn leads to more accurate estimates of focal group mean theta and hence lower RMSE values.

Discussion

To our knowledge, this is the first simulation study to establish that the IRT-C methodology can be used to ascertain DIF. We examine several variations of the IRT-C procedure to determine which best controls for the Type I error rates while giving reasonable power. In view of this, we determined the power and Type I error rate for detecting DIF with the IRT-C using either UBVR or CBVR. We found that CBVR generally performed equally well or better than UBVR across a variety of DIF conditions. Further, non-uniform DIF is better detected with CBVR, and the power and Type I error rates are better for non-compensatory DIF conditions, especially when the proportion of DIF items in a scale is large. Different stopping criteria for the IRT-C iterative procedure were also examined. Using AIC3 as a stopping criterion produced results that were equivalent to not using any information criteria and hence we do not advocate its use. The primary difference between the using no information criteria and the BIC criterion is the degree to which the Type I error rates are controlled for: using the BIC stopping rule leads to a more conservative Type I error rate whereas the not using information criteria leads to better control of the Type I error rate.

The IRT-C procedure performed relatively well compared to the two other established DIF techniques – the Mantel-Haenszel and the MIMIC procedure. When there is a small proportion of items on the scale that have DIF, the IRT-C procedure without the use of information criteria performs the best in that the Type I error rates are well controlled at 0.05 across all the DIF conditions simulated and it has the highest power, especially to detect non-uniform DIF. However, when a large proportion of items on the scale have DIF, the IRT-C procedure does not control well for the Type I error rates in some occasions resulting in an averaged Type I error rate of around 0.10 whereas the MIMIC procedure had good control of the Type I error rates at 0.05 across the simulated DIF conditions.

Why does the IRT-C procedure have less control over the Type I error rates (compared to the MIMIC procedure) in the case of a large proportion of items exhibiting DIF? One reason

is that the MIMIC procedure using a free baseline approach in which the designated anchor item is simulated to be invariant across groups. This ensures that the metrics are correctly linked across groups. However, for the IRT-C procedure, it uses a constrained baseline approach without any designated anchors. When 40 percent of the items are consistently biased in the same direction, the metrics are not accurately linked. Therefore, when there is non-compensatory DIF, the IRT-C procedure does not fair as well; however, when there is compensatory DIF, the IRT-C procedure still performs reasonably well and often has higher power than the MIMIC procedure.

One advantage of the IRT-C procedure as currently proposed is that one does not need to have a predetermined set of invariant anchors. In practice, the appropriate set of anchors may sometimes be difficult to ascertain and the incorrect set of anchors (on which some may have DIF) could result in less control of Type I error rates. Further, a separate procedure is necessary to pre-identify anchor items before undertaking the IRT-DIF MIMIC procedure (see Woods, 2009b).

Limitations and Future Research

Although past research on DIF usually focuses on the detection of DIF in general, one advantage of the IRT-C procedure is detecting specific forms of DIF (i.e., uniform or non-uniform). This is an important direction because recent theory and research suggests that the causes of uniform and non-uniform DIF on multiple-choice tests are distinct (Penfield, 2010). Although the IRT-C procedure detects uniform DIF well, it does not detect non-uniform DIF as well. Therefore, more research is needed to determine how to improve power for detecting non-uniform DIF with the IRT-C procedure. Interestingly, it appears the mere detection of DIF in items, rather than the detection of specific forms of DIF, is sufficient to ensure an accurate estimation of the latent mean difference between groups. However, future research can examine

whether predictive validities would be affected to the degree that non-uniform DIF is undetected.

Both the IRT-C and the MIMIC procedures allow the use of multiple covariates and multiple categories. This is an advantage over traditional DIF statistics in which we are limited to a single categorical variable (i.e., two-group or multiple group comparisons). With IRT-C and the MIMIC model, it is possible to ascertain whether DIF occurs in one variable (e.g., age) over another variable (e.g., gender). Although this study and past research suggests that for a dichotomous covariate (i.e., two-group case), the IRT-C and the MIMIC model works well, more research needs to examine whether DIF can be detected in the case where multiple covariates are used.

One of the first few studies examining the MIMIC model for DIF assessment began with the use of dichotomous data (Woods, 2009a, 2009b). Similarly, because this is the first study that seeks to validate the IRT-C procedure, we chose to focus on dichotomous data with a range of DIF conditions. We sought to determine how well uniform or nonuniform DIF were detected, the stopping rule that would work best, and the type of BVR that is most sensitive to DIF. Given that the IRT-C procedure performs well with dichotomous data, more research needs to examine whether the results would generalize to polytomous data.

More research can also examine whether the IRT-C can be applied to a testing context. For instance, the 3-parameter logistic model (3PLM) is commonly used in large scale educational testing. To our knowledge, one advantage of the IRT-C method is that it can model 3PLM responses, unlike the MIMIC method. Recent research shows that the 3PLM IRT-C procedure effectively detects DIF and accurately recovers the latent means of different covariate groups (Tay, Drasgow, & Vermunt, 2011).

The IRT-C procedure is based on the null hypothesis statistical testing paradigm. This can be a limiting particularly when the sample size is large as many items can be flagged for

DIF despite minuscule DIF effect sizes. Although the null hypothesis statistical testing paradigm to examining DIF is widely used in many traditional statistical testing methods, we propose that more research can focus on procedures that incorporate the evaluation of DIF effect sizes in addition to statistical testing (cf. Steinberg & Thissen, 1996).

Conclusion

To our knowledge, this is the first IRT analog to the MIMIC procedure for detecting DIF. This study has shown that the IRT-C procedure can be as, or even more effective than the Mantel-Haenzsel or MIMIC procedures for detecting DIF. One advantage of this approach is that anchor items do not need to be known *a priori*. We advocate the use of the IRT-C method and encourage more research in this area.

References

Bartholomew, D. J., & Tzamourani, P. (1999). The goodness-of-fit of latent trait models in attitude measurement. *Sociological Methods & Research, 27*, 525-546.

Bolt, D. M. (2005). Limited- and full-information estimation of item response theory models. In D. M. Bolt & J. J. McArdle (Eds.), *Contemporary psychometrics: a festschrift for Roderick P. McDonald*. Mahwah, New Jersey: Lawrence Erlbaum.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2^P tables. *British Journal of Mathematical and Statistical Psychology, 59*, 173-194.

Dias, J. M. G. (2004). *Finite mixture models: Review, applications and computer intensitve methods.* University of Groningen, The Netherlands.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B. A., & Mead, A. D. (1995). Fitting polytomous item response models to multiple-choice tests. *Applied Psychological Measurement, 19*, 145-165.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*, 278-295.

Finch, H., & French, B. F. (2008). Anomalous Type I error rates for identifying one type of differential item functioning in the presence of the other. *Educational and Psychological Measurement, 68*(742-759).

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange Multiplier tests. *Statistica Sinica, 8*, 647-667.

Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika, 64*, 273-294.

Holland, W. P., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

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.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Matsumoto, D., & Yoo, S. H. (2006). Toward a new generation of cross-cultural research. *Perspectives on psychological science, 1*(3), 234-250. doi: 10.1111/j.1745-6916.2006.00014.x

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7*, 361-388. doi: 10.1177/1094428104268027

Meade, A. W., Lautenschlager, G. J., & Johnson, E. C. (2007). A Monte Carlo examination of the sensitivity of the differential functioning of items and tests framework for tests of measurement invariance with Likert data. *Applied Psychological Measurement, 31*, 430-455.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297-334.

Muthén, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics, 10*(121-132).

Muthén, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 213-238).

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 1994*, 315-328.

Paek, I. (2010). Conservativeness in rejection of the null hypothesis when using the continuity correction in the MH chi-square test in DIF applications. *Applied Psychological Measurement, 34*, 539-548.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education, 2001*, 235-259.

Penfield, R. D. (2010). Modeling DIF effects using distractor-level invariance effects: Implications for understanding the causes of DIF. *Applied Psychological Measurement, 34*, 151-165. doi: 10.1177/0146621609359284

R, D., Core, Team. (2008). R: A language and environment for statistical computing. Vienna, Austria: ISBN 3-900051-07-0, URL http://www.R-project.org.

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

Rogers, J. H., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105-116.

Saris, W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. *Sociological Methodology, 17,* 105-129.

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

Shih, C.-L., & Wang, W.-C. (2009). Differential item functioning detecting using Multiple Indicators, Multiple Causes method with a pure short anchor. *Applied Psychological Measurement, 33*, 184-199.

Sörbom, D. (1989). Model modification. *Psychometrika, 54*, 371-384.

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.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11*, 402–415.

Stone, C. A. (1992). Recovery of marginal maximum likellihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*, 1-16.

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

Tay, L., Ali, U. S., Drasgow, F., & Williams, B. A. (2011). Fitting simulated dichotomous and polytomous data: Examining the difference between ideal point and dominance models. *Applied Psychological Measurement, 35*, 280-295.

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*.

Tay, L., Drasgow, F., & Vermunt, J. K. (2011). *An IRT approach for simultaneously assessing the relative impact of different variables on SAT® latent scores and differential item functioning.* Technical Report, New York: College Board.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). New Jersey: Hillsdale.

Van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47*, 123-140.

Vermunt, J. K., & Magidson, J. (2000). *Latent GOLD User's Guide*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for Latent GOLD 4.0: Basic and advanced*. Belmont Massachusetts: Statistical Innovations Inc.

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

Vermunt, J. K., van Ginkel, J. R., van der Ark, L. A., & Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology, 33*, 369-397.

Wang, W.-C., Shih, C.-L., & Yang, C.-C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement, 69*, 713-731.

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

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

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.

Table 1

Comparison of DIF procedures: 20% DIF and 250 focal group members

				IRT-C								Mantel-Haenszel		MIMIC	
				No information criteria				BIC stopping rule							
					Power.		RMSE		Power.		RMSE				
Condition	***C*[a]**	***a*[b]**	***b*[c]**	**Power**	**Form**	**T1E**	**$\theta_{f.mean}$**	**Power**	**Form**	**T1E**	**$\theta_{f.mean}$**	**Power**	**T1E**	**Power**	**T1E**
1	-	-	-	-	-	0.05	0.09	-	-	0.01	0.09	-	0.04	-	0.05
2	N	-0.4	0	0.55	0.39	0.05	0.10	0.35	0.23	0.01	0.10	0.27	0.07	0.25	0.05
3	Y	-0.4	0	0.55	0.35	0.04	0.10	0.32	0.19	0.01	0.09	0.36	0.05	0.26	0.04
4	N	0	0.4	0.67	0.64	0.05	0.11	0.45	0.44	0.01	0.11	0.64	0.08	0.49	0.06
5	Y	0	0.4	0.71	0.69	0.04	0.10	0.53	0.50	0.01	0.10	0.81	0.05	0.57	0.06
6	N	-0.4	0.4	0.57	0.09	0.05	0.09	0.32	0.07	0.01	0.09	0.50	0.08	0.52	0.07
7	Y	-0.4	0.4	0.67	0.13	0.05	0.10	0.46	0.11	0.01	0.10	0.62	0.05	0.49	0.07
8	N	0	0.8	0.96	0.92	0.05	0.09	0.94	0.90	0.01	0.09	0.92	0.19	0.92	0.05
9	Y	0	0.8	0.98	0.95	0.04	0.10	0.98	0.93	0.01	0.10	0.97	0.06	0.93	0.04
10	N	-0.4	0.8	0.75	0.30	0.05	0.09	0.52	0.19	0.02	0.10	0.89	0.21	0.90	0.08
11	Y	-0.4	0.8	0.82	0.24	0.05	0.10	0.74	0.21	0.01	0.10	0.89	0.07	0.83	0.05
				0.72	**0.47**	**0.05**	**0.10**	**0.56**	**0.38**	**0.01**	**0.10**	**0.69**	**0.09**	**0.61**	**0.06**

Note. For comparison, no DIF was simulated in Condition 1; [a] Was compensatory DIF simulated? [b] Degree of *a*-parameter DIF simulated for focal group; [c] Degree of *b*-parameter DIF for focal group; Power represents the proportion of items in which DIF was significant averaged across the replications; Power.Form represents power to detect the exact form of DIF simulated; T1E represents the Type I error rates; RMSE $\theta_{f.mean}$ represents the root mean squared error of the estimated focal latent mean against that of the simulated focal latent mean.

Table 2

Comparison of DIF procedures: 20% DIF and 500 focal group members

				IRT-C								Mantel-Haenszel		MIMIC	
				No information criteria				BIC stopping rule							
					Power.		RMSE		Power.		RMSE				
Condition	***C*[a]**	***a*[b]**	***b*[c]**	**Power**	**Form**	**T1E**	**$\theta_{f.mean}$**	**Power**	**Form**	**T1E**	**$\theta_{f.mean}$**	**Power**	**T1E**	**Power**	**T1E**
1	-	-	-	-	-	0.05	0.08	-	-	0.01	0.07	-	0.05	-	0.07
2	N	-0.4	0	0.67	0.49	0.06	0.08	0.49	0.37	0.01	0.08	0.36	0.09	0.33	0.04
3	Y	-0.4	0	0.62	0.44	0.06	0.09	0.45	0.33	0.01	0.08	0.45	0.06	0.33	0.05
4	N	0	0.4	0.80	0.77	0.05	0.09	0.64	0.62	0.01	0.08	0.80	0.11	0.68	0.05
5	Y	0	0.4	0.84	0.80	0.06	0.08	0.74	0.72	0.01	0.08	0.92	0.05	0.69	0.05
6	N	-0.4	0.4	0.68	0.17	0.06	0.08	0.57	0.15	0.01	0.07	0.68	0.13	0.71	0.06
7	Y	-0.4	0.4	0.72	0.22	0.06	0.08	0.67	0.24	0.01	0.07	0.69	0.06	0.62	0.06
8	N	0	0.8	1.00	0.95	0.04	0.08	0.98	0.92	0.01	0.08	0.97	0.31	0.97	0.05
9	Y	0	0.8	1.00	0.95	0.04	0.08	1.00	0.94	0.01	0.08	0.99	0.08	0.98	0.04
10	N	-0.4	0.8	0.81	0.39	0.05	0.08	0.73	0.33	0.01	0.08	0.95	0.32	0.95	0.06
11	Y	-0.4	0.8	0.90	0.39	0.05	0.07	0.82	0.39	0.00	0.08	0.94	0.07	0.88	0.04
				0.80	**0.55**	**0.05**	**0.08**	**0.71**	**0.50**	**0.01**	**0.08**	**0.77**	**0.12**	**0.71**	**0.05**

Note. For comparison, no DIF was simulated in Condition 1; [a] Was compensatory DIF simulated? [b] Degree of *a*-parameter DIF simulated for focal group; [c] Degree of *b*-parameter DIF for focal group; Power represents the proportion of items in which DIF was significant averaged across the replications; Power.Form represents power to detect the exact form of DIF simulated; T1E represents the Type I error rates; RMSE $\theta_{f.mean}$ represents the root mean squared error of the estimated focal latent mean against that of the simulated focal latent mean.

Table 3

Comparison of DIF procedures: 40% DIF and 250 focal group members

				IRT-C								Mantel-Haenszel		MIMIC	
				No information criteria				BIC stopping rule							
					Power.		RMSE		Power.		RMSE				
Condition	***C*[a]**	***a*[b]**	***b*[c]**	**Power**	**Form**	**T1E**	**$\theta_{f.mean}$**	**Power**	**Form**	**T1E**	**$\theta_{f.mean}$**	**Power**	**T1E**	**Power**	**T1E**
1	-	-	-	-	-	0.05	0.09	-	-	0.01	0.09	-	0.04	-	0.09
2	N	-0.4	0	0.36	0.22	0.09	0.11	0.15	0.09	0.03	0.11	0.13	0.12	0.21	0.05
3	Y	-0.4	0	0.44	0.25	0.07	0.12	0.29	0.18	0.01	0.10	0.33	0.07	0.24	0.05
4	N	0	0.4	0.39	0.36	0.15	0.21	0.24	0.23	0.07	0.19	0.40	0.20	0.52	0.05
5	Y	0	0.4	0.70	0.67	0.08	0.12	0.52	0.50	0.02	0.13	0.80	0.06	0.53	0.05
6	N	-0.4	0.4	0.39	0.05	0.12	0.11	0.18	0.03	0.04	0.11	0.34	0.21	0.53	0.07
7	Y	-0.4	0.4	0.63	0.16	0.07	0.11	0.46	0.13	0.02	0.12	0.59	0.06	0.49	0.06
8	N	0	0.8	0.76	0.73	0.23	0.34	0.68	0.65	0.20	0.37	0.81	0.58	0.94	0.04
9	Y	0	0.8	0.97	0.92	0.07	0.13	0.95	0.91	0.03	0.14	0.97	0.09	0.93	0.05
10	N	-0.4	0.8	0.52	0.13	0.13	0.15	0.32	0.09	0.08	0.14	0.80	0.56	0.89	0.07
11	Y	-0.4	0.8	0.81	0.27	0.08	0.13	0.72	0.24	0.03	0.12	0.91	0.09	0.81	0.07
				0.60	**0.38**	**0.10**	**0.15**	**0.45**	**0.30**	**0.05**	**0.15**	**0.61**	**0.19**	**0.61**	**0.06**

Note. For comparison, no DIF was simulated in Condition 1; [a] Was compensatory DIF simulated? [b] Degree of *a*-parameter DIF simulated for focal group; [c] Degree of *b*-parameter DIF for focal group; Power represents the proportion of items in which DIF was significant averaged across the replications; Power.Form represents power to detect the exact form of DIF simulated; T1E represents the Type I error rates; RMSE $\theta_{f.mean}$ represents the root mean squared error of the estimated focal latent mean against that of the simulated focal latent mean.

Table 4

Comparison of DIF procedures: 40% DIF and 500 focal group members

				IRT-C											
				No information criteria				BIC stopping rule				Mantel-Haenszel		MIMIC	
					Power.		RMSE		Power.		RMSE				
Condition	C^a	a^b	b^c	**Power**	**Form**	**T1E**	$\theta_{f.mean}$	**Power**	**Form**	**T1E**	$\theta_{f.mean}$	**Power**	**T1E**	**Power**	**T1E**
1	-	-	-	-	-	0.06	0.08	-	-	0.01	0.07	-	0.04	-	0.09
2	N	-0.4	0	0.49	0.31	0.12	0.11	0.27	0.20	0.03	0.10	0.22	0.17	0.28	0.06
3	Y	-0.4	0	0.55	0.37	0.08	0.11	0.39	0.29	0.02	0.10	0.43	0.07	0.33	0.04
4	N	0	0.4	0.61	0.57	0.15	0.17	0.38	0.36	0.07	0.18	0.58	0.29	0.67	0.06
5	Y	0	0.4	0.81	0.78	0.07	0.10	0.66	0.63	0.01	0.11	0.91	0.06	0.69	0.06
6	N	-0.4	0.4	0.54	0.09	0.14	0.10	0.30	0.06	0.04	0.09	0.51	0.32	0.66	0.08
7	Y	-0.4	0.4	0.67	0.27	0.06	0.09	0.56	0.23	0.03	0.10	0.70	0.09	0.56	0.05
8	N	0	0.8	0.80	0.76	0.25	0.37	0.79	0.75	0.19	0.34	0.89	0.74	0.97	0.03
9	Y	0	0.8	0.98	0.92	0.06	0.12	0.98	0.91	0.03	0.14	0.99	0.12	0.98	0.05
10	N	-0.4	0.8	0.73	0.25	0.11	0.12	0.54	0.21	0.08	0.14	0.90	0.69	0.96	0.06
11	Y	-0.4	0.8	0.88	0.46	0.05	0.09	0.82	0.45	0.03	0.09	0.94	0.12	0.89	0.05
				0.71	**0.48**	**0.11**	**0.13**	**0.57**	**0.41**	**0.05**	**0.13**	**0.71**	**0.25**	**0.70**	**0.06**

Note. For comparison, no DIF was simulated in Condition 1; [a] Was compensatory DIF simulated? [b] Degree of *a*-parameter DIF simulated for focal group; [c] Degree of *b*-parameter DIF for focal group; Power represents the proportion of items in which DIF was significant averaged across the replications; Power.Form represents power to detect the exact form of DIF simulated; T1E represents the Type I error rates; RMSE $\theta_{f.mean}$ represents the root mean squared error of the estimated focal latent mean against that of the simulated focal latent mean.

Appendix

Bivariate Residuals

Differential item functioning (DIF) can be determined by the use of a fully constrained baseline approach (see Stark, Chernyshenko, & Drasgow, 2006). Using this constrained baseline model, it was proposed that large unconditional bivariate residuals (UBVR) between covariates (or grouping variables) and items are an indication of DIF (Tay, Newman & Vermunt, 2011). For values with large UBVRs, the parameters associated with DIF (i.e., c_i or d_i in Equation 2) are then estimated and tested for significance using the Wald statistic. For example, the table below shows the second item may have DIF on the first covariate (z_1) because it has the largest UBVR value. Subsequently, the parameters c_2 or d_2 for z_1 are freely estimated and tested for significance.

	Item (i)				
Covariate (z)	i = 1	i = 2	i =3	…	i = 10
z_1	1.01	**4.52**	3.65		1.41
z_2	2.61	3.12	3.87		1.59

The UBVR implemented in Latent GOLD 4.5 (LG; Vermunt & Magidson, 2008) is a Pearson χ^2-type measure reflecting the discrepancy between observed and expected cell counts in the two-way table cross-tabulating a covariate (or an observed grouping variable) and an item. In fact, it quantifies the *z*-*y* association that is not explained by the specified IRT model. As indicated by Vermunt and Magidson (2005), an UBVR is a lower bound for the reduction of the -2 log-likelihood when including the c_i term (see Equation 2) into the IRT model. This is an approximation for the Lagrange multiplier test described by Glas (1998; 1999), and is similar to modification indices used in structural equation modeling (Saris, Satorra, & Sörbom, 1987; Sörbom, 1989). It should be noted that the use of χ^2 statistics from marginal two-way tables has not only been proposed for model modification, but also for goodness-of-fit testing of IRT models (Bartholomew & Tzamourani, 1999; Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Drasgow, Levine, Tsien, Williams, & Mead, 1995).

For each covariate, the formulation of the UBVR is as follows. Let $n(y_i = r, z = g)$ and $m(y_i = r, z = g)$ be the observed and expected cell entries in the table cross-tabulating the grouping variable and item *i*, where *g* and r refer to a particular group and response option for item *i*, respectively. Assuming that there are *G* groups and that the items have *R* response categories, the UBVR for item *i* is defined as follows:

$$BVR_{uncondional,i} = \sum_{r=1}^{R}\sum_{g=1}^{G} \frac{\left[n(y_i = r, z = g) - m(y_i = r, z = g)\right]^2}{m(y_i = r, z = g)} \quad . \tag{5}$$

The expected cells entries in relevant two-way tables are defined as follows:

$$m(y_i = r, z = g) = \sum_{j=1}^{J} \int P(y_{ji} = r \mid \theta)\ f(\theta \mid z_j = g, \mathbf{y}_j)\, d\theta, \tag{6}$$

that is, by integrating over the latent trait θ. However, because no close from expression exists for the integral, this is solved numerically using *K* quadrature nodes yielding the following approximation

$$m(y_i = r, z = g) = \sum_{j=1}^{J}\sum_{k=1}^{K} P(y_{ji} = r \mid \theta_k)\ \pi(\theta_k \mid z_j = g, \mathbf{y}_j). \quad (7)$$

The term $\pi(\theta_k \mid z_j = g, \mathbf{y}_j)$ is the posterior probability associated with the k[th] quadrature node. It is the same probability as needed in the computation of the derivatives required for the marginal maximum likelihood estimation of the IRT-C model.

The observed cell counts are obtained by counting the number of persons with $y_i = r$ and $z = g$, which can be expressed as follows:

$$n(y_i = r, z = g) = \sum_{j=1}^{J} I(y_{ji} = r, z_j = g) \quad (8)$$

where $I(\bullet)$ equal 1 if the expression is true and 0 otherwise. BVR values can be expected to be inflated when a single set of item parameters does not fully account for the observed cell entries, which may be an indication of DIF.

In the computation of the UBVR, one integrates out the latent trait before computing the discrepancy between observed and expected frequencies. A disadvantage of such a procedure is that information on non-uniform DIF (on the fact that the trait-response relationship may differ across groups) gets lost (cf. Van den Wollenberg, 1982). To deal with this issue, we propose and examine an alternative discrepancy measure – the conditional BVR (CBVR) -- that is implemented as follows:

$$BVR_{condional,i} = \sum_{k=1}^{K}\sum_{r=1}^{R}\sum_{g=1}^{G} \frac{\left[n(y_i = r, z = g \mid \theta_k) - m(y_i = r, z = g \mid \theta_k)\right]^2}{m(y_i = r, z = g \mid \theta_k)} \quad (9)$$

where

$$m(y_i = r, z = g \mid \theta_k) = \sum_{j=1}^{J} P(y_{ji} = r \mid \theta_k)\ \pi(\theta_k \mid z_j = g, \mathbf{y}_j) \quad (10)$$

and

$$n(y_i = r, z = g \mid \theta_k) = \sum_{j=1}^{J} I(y_{ji} = r, z_j = g \mid \theta_k)\pi(\theta_k \mid z_j = g, \mathbf{y}_j) \quad (11)$$

Note that $m(y_i = r, z = g \mid \theta_k)$ is in fact an expected cell count in the three-way cross-table of group, item *i* and a node of integration, and $n(y_i = r, z = g \mid \theta_k)$ is an estimate of the corresponding observed cell entry. This proposed residual has recently been implemented in Latent GOLD 5.0.