Using Hierarchical Logistic Regression to Study DIF and DIF Variance in Multilevel Data

Benjamin R. Shear

University of Colorado Boulder

October, 2018

Pre-print version of:

Shear, B. R. (2018). Using hierarchical logistic regression to study DIF and DIF variance in multilevel data. *Journal of Educational Measurement*, *55*(4), 513–542. https://doi.org/10.1111/jedm.12190

Using Hierarchical Logistic Regression to Study DIF and DIF Variance in Multilevel Data

Benjamin R. Shear[1]

[1] University of Colorado Boulder

Author Note

The author would like to thank Ed Haertel, sean reardon, Ben Domingue, Bruno Zumbo, Derek Briggs, and several anonymous reviewers for helpful comments on this work. A portion of this work was completed as part of the author's doctoral dissertation and supported in part by an Institute for Education Sciences Pre-Doctoral Training Grant (#R305B090016).

Correspondence concerning this article should be addressed to Benjamin R. Shear, 249 UCB, Boulder, CO 80309. E-mail: benjamin.shear@colorado.edu

Abstract

When contextual features of test-taking environments differentially affect item responding for different test-takers and these features vary across test administrations, they may cause differential item functioning (DIF) that varies across test administrations. Because many common DIF detection methods ignore potential DIF variance, this paper proposes the use of random coefficient hierarchical logistic regression (RC-HLR) models to test for both uniform DIF and DIF variance simultaneously. A simulation study and real data analysis are used to demonstrate and evaluate the proposed RC-HLR model. Results show the RC-HLR model can detect uniform DIF and DIF variance more accurately than standard logistic regression DIF models in terms of bias and Type I error rates.

Keywords: differential item functioning; validity; multilevel models; logistic regression

Using Hierarchical Logistic Regression to Study DIF and DIF Variance in Multilevel Data

Differential item functioning (DIF; Holland & Thayer, 1988; Holland & Wainer, 1993) is a psychometric framework for studying potential item and test bias based on patterns in item responses. DIF analyses are widely used both in the test development and validation processes. An item is said to show DIF if there are systematic differences in the likelihood of answering an item correctly between two or more groups of test-takers, after matching the groups on the construct being measured by the test. While DIF can be analyzed for tests measuring a wide range of constructs, the construct being measured is often referred to as "ability" for shorthand. DIF is problematic because, depending upon the cause of the DIF, it may invalidate proposed test score interpretations for some test-takers. During test development, DIF analyses can be used to flag problematic items and to test assumptions of parameter invariance necessary for the use of item response theory for test scaling and equating (Hambleton, Swaminathan, & Rogers, 1991). More recently, Zumbo et al. (2015) described how DIF analyses can be used to study features of the test-taking context that affect item responding, and hence contribute to our understanding of score meaning. In all cases, DIF analyses provide an important source of validity evidence both to ensure appropriate test score inferences and test fairness (AERA, APA, & NCME, 2014).

Although large-scale tests are intended to be administered under standardized conditions, certain test-taking conditions can vary across specific administrations of a test. Here the term test administration refers to a single instance in which a group of test-takers completes a test form that will also be taken by other test-takers at other times or locations. Features of the testing situation in any specific administration of the test, such as the size of the testing room, the instructions provided by the test proctor, the demographic make-up of other test-takers, or test preparation activities can vary, even when the test content and format remain constant. When item response data from multiple administrations of a test form are collected, there is a hierarchical or multilevel structure to the data, with test-takers nested within test administrations. The ecological model of item responding described by Zumbo et al. (2015) highlights the fact that features of the test administration environment, and other ecological variables, could systematically affect item responses.

If the source of DIF for a particular test item depends on aspects of the testing situation that vary across administrations, then many common DIF methods may prove inadequate for detecting and understanding the resulting DIF. Stereotype threat (Steele, 1997), for example, provides one theoretical framework for understanding how and why DIF might vary across test administrations. Stereotype threat refers to "the social-psychological threat that arises when one is in a situation or doing something for which a negative stereotype about one's group applies" (Steele, 1997, p. 614). Prior studies have documented that aspects of the testing situation such as framing a test as evaluative of one's ability (Steele & Aronson, 1995; Steele, Spencer, & Aronson, 2002) or varying the demographic make-up of other test-takers (Inzlicht & Ben-Zeev, 2000, 2003) can induce stereotype threat and subsequently reduce test performance for stereotyped groups. Stereotype threat is thus a phenomenon that could, in theory, cause DIF between stereotyped and non-stereotyped groups during some test administrations, but not others. When an item displays DIF that varies across test administrations, it will be referred to here as "DIF variance."

Applying models that ignore DIF variance could be problematic for a number of reasons. First, a model estimating a single DIF statistic will be mis-specified, potentially resulting in biased parameter estimates or incorrect statistical significance tests. Second, ignoring heterogeneity in DIF across test administrations could mask important patterns in item responses, making it difficult to identify the true source of the DIF, a commonly encountered problem in practice (Angoff, 1993). Third, framing DIF as a phenomenon that potentially varies across test administrations highlights the multilevel nature of most test score data; even if there is no DIF variance, the multilevel nature of the data can still be modeled appropriately and should lead to more accurately estimated standard errors (e.g., French & Finch, 2010; Jin, Myers, & Ahn, 2014).

To address these concerns, this paper describes how hierarchical logistic regression (HLR; Hox, 2010) models can be used to quantify and study DIF variance. The next section provides background on the concept of DIF variance and the subsequent sections describe how HLR models can be used to extend the standard logistic regression (LR) DIF models to study DIF variance. A computer simulation evaluating the efficacy of the HLR model for detecting DIF variance is then described, and an illustrative real data example demonstrates how the HLR model can be applied to study DIF between male and female students taking the same mathematics test at different schools. The final sections summarize results and discuss potential directions for future work.

Background

Historically, DIF detection methods have been used primarily for test and item development, and as a reaction to public concerns over test bias such as the widely discussed Golden Rule case (Anrig, 1987). Zumbo (2007) describes three broad generations in the development and use of DIF methodology. The first generation focused on studying “item bias” and was characterized by introducing the initial conceptual distinctions between item bias, item impact, and DIF. The second generation was characterized by acceptance of these terms and development of statistical methods to identify DIF, including contingency table methods such as the Mantel-Haenszel method (Holland & Thayer, 1988), standardized difference method (Dorans & Kulick, 1986), logistic regression (Swaminathan & Rogers, 1990), latent variable models such as item response theory (Meredith, 1993; Millsap, 2011), and multidimensionality models such as SIBTEST (Shealy & Stout, 1993).

The third (and current) generation is characterized by an expanding array of purposes for which DIF analyses might be used, such as better understanding examinees’ item response processes (Zumbo, 2007). Drawing on the work of Bronfenbrenner and others, Zumbo et al. (2015) and Zumbo and Gelin (2005) describe an ecological model of item responding intended to help guide third generation DIF analyses. This ecological model of item responding seeks to orient researchers' focus towards "sociological, structural, community, and contextual variables, as well as psychological and cognitive factors, as explanatory sources of item responding and hence of DIF" (Zumbo et al., 2015, p. 139). The ecological model of item responding acknowledges the possibility that factors in the testing situation might systematically affect individual test-takers' item responses, and that these effects may differ for different groups of test-takers.

The third generation of DIF can be "characterized as conceiving of DIF as occurring because of some characteristic of the test item *and/or testing situation* that is not relevant to the underlying ability of interest (and hence the test purpose)" (Zumbo, 2007, p. 229, emphasis original). The key turn here is a consideration of the testing situation as a potential source or moderator of DIF, in addition to consideration of item and test-taker characteristics. Millsap (2011) makes a similar point, noting that a focus only on individual differences as sources of bias may "blind investigators to other explanations, such as those that are environmental, situational, or social in origin" (p. 55). Similar to the example of stereotype threat above, Millsap provides the example of a researcher administering a measure of racial prejudice, noting that the race of the researcher administering the survey may affect participants' responses and the effects could differ across different populations of participants. As another example, researchers often test for DIF between English language learners (ELL) and native English speakers; if some schools coach ELL students on strategies for interpreting complex or ambiguous wordings on test items, DIF variance could potentially arise across schools if this coaching reduces or eliminates DIF.

In these examples, DIF is affected by one or more features of the testing situation and can be conceptualized as a form of moderated DIF (Zumbo et al., 2015). When DIF is moderated by features of the testing situation that vary across administrations, this can lead to DIF variance. Identifying features of a testing situation that can explain variance in DIF could lead to a better understanding of the mechanism of the DIF, and help to more fully explain observed variation in item responses. The presence of DIF variance provides evidence that a feature of the testing situation may be moderating the observed DIF, although it could also indicate a case in which the DIF is mediated by a third variable that varies systematically across test administrations. This paper focuses on a method for identifying the presence of DIF variance, which would be conducted prior to studies attempting to identify variables that can explain the variance in the DIF. More specifically, the next two sections describe how the LR DIF detection framework can be extended to quantify and test for potential DIF variance. The proposed method is then tested with a simulation and illustrated with a real data analysis.

Logistic Regression DIF Model

LR provides an effective method for DIF detection and has the practical advantage that, as a member of the generalized linear model family (Agresti, 2013), it can be extended in numerous ways (Zumbo, 1999). To introduce the LR DIF model, assume a test consists of items scored dichotomously as 1 if answered correctly and 0 otherwise. The LR DIF model tests whether item performance for one group of test-takers, usually referred to as the "focal" group (often a particular racial or ethnic minority group, or another relevant group of interest), differs statistically from that of a "reference" group of test-takers, after matching the two groups on ability. The LR model for detecting DIF in a single item can be formulated as (Swaminathan & Rogers, 1990; Zumbo, 1999):

$$\ln\left(\frac{Pr[Y_i = 1]}{1 - Pr[Y_i = 1]}\right) = \beta_0 + \beta_1 x_i + \beta_2 g_i + \beta_3 (x_i \times g_i), \tag{1}$$

where ln() is the natural logarithm, $Pr[Y_i = 1]$ is the probability that examinee i correctly responds to the target item, x_i is a measure of ability for examinee i, and g_i is a binary indicator equal to 1 if examinee i is a member of the focal group and 0 if a member of the reference group. In a DIF analysis, this model is fit separately for each item. Item subscripts are removed for clarity of notation. The matching variable x_i is most often a total score calculated as the sum of the number of items answered correctly, including the studied item.

In Equation 1, β_1 is similar to an item discrimination parameter, while β_2 represents a constant difference in item difficulty (i.e., the log-odds of a correct response) between reference and focal groups, after conditioning on ability. A non-zero value of β_2 indicates the presence of "uniform" DIF because the difficulty of the item differs across matched groups by a constant (i.e., "uniform") amount. A nonzero value of β_3 indicates that the difference in item difficulty between groups changes linearly across the ability distribution and is often referred to as "nonuniform" DIF (Swaminathan & Rogers, 1990). Note that although a non-zero β_3 term implies heterogeneity of DIF across the ability distribution, the effect is assumed to be fixed across test administrations and hence it does not imply the form of heterogeneity across test administrations that would be due to DIF variance.

In this paper I consider extensions of the reduced LR model used to test for uniform DIF that excludes the β_3 term:

$$\ln\left(\frac{Pr[Y_i = 1]}{1 - Pr[Y_i = 1]}\right) = \beta_0 + \beta_1 x_i + \beta_2 g_i. \tag{2}$$

Testing for DIF proceeds by estimating the parameters of this model for each item separately (usually using maximum likelihood estimation). Items with statistically significant β_2 coefficients are flagged for DIF.[1] In practice, DIF is often assessed using the full model in Equation 1. The nonuniform DIF term is excluded from the model here to simplify the current exposition and for consistency with prior studies using HLR DIF models, but the framework could be extended to incorporate the nonuniform DIF term in future research.

Hierarchical Logistic Regression DIF Model

When there is a relevant multilevel structure in the data, HLR DIF models that explicitly incorporate this multilevel structure are recommended (French & Finch, 2010). Administering a common test form to test-takers at different times or locations creates a multilevel structure in the data. In the case of state achievement testing, for example, the same statewide achievement test is administered separately at each school. Each school would be considered a separate test administration. In the multilevel modeling framework, these groupings are often referred to as "clusters;" in this case, test-takers would be the level 1 units "nested" within test administrations or "clusters." Here I assume that the clusters are observed groupings in the data set and that the same test form was taken by test-takers across all clusters. The appropriate level of aggregation could vary depending upon the research questions and hypotheses. For example if one believes that relevant features of the testing situation vary across the actual rooms in which tests are administered, then each room could represent a cluster. On the other hand, if one believes there are factors at the level of the school building (or higher levels of aggregation) that moderate DIF for all students within that unit, then each of these higher level units would represent a cluster. In the remainder of the paper, the terms test administration and cluster will be used interchangeably to refer to the groupings in the dataset across which DIF varies.

Random Coefficient HLR Model

To quantify potential DIF variance, let there be a set of $j = 1, 2, \ldots, J$ clusters representing J different administrations of the same test. The LR DIF model in Equation 2 can be extended to a 2-level HLR model for each item (again, item subscripts are not shown) with level 1 equation

$$\ln\left(\frac{Pr[Y_{ij}=1]}{1-Pr[Y_{ij}=1]}\right)=\beta_{0j}+\beta_{1j}x_{ij}+\beta_{2j}g_{ij}, \tag{3}$$

where $Pr[Y_{ij}=1]$ is the probability that student i in cluster j responds correctly to the studied item, and level 2 equations

$$\begin{aligned}\beta_{0j}&=\gamma_{00}+u_{0j}\\ \beta_{1j}&=\gamma_{10}+u_{1j}\\ \beta_{2j}&=\gamma_{20}+u_{2j},\end{aligned} \tag{4}$$

where the $u_{\bullet j}$ are multivariate normal random effects with mean 0 and covariance matrix

$$\mathbf{T}=\begin{bmatrix}\tau_{00} & & \\ \tau_{10} & \tau_{11} & \\ \tau_{20} & \tau_{21} & \tau_{22}\end{bmatrix}. \tag{5}$$

This is the same DIF model presented in Equation 2, but the coefficients are modeled as random variables that are normally distributed across clusters. This is the most general form of what will be referred to here as the random coefficient HLR (RC-HLR) model. In the simulations and analyses below, further constraints are placed on the **T** matrix.

The γ_{10} term represents the average item discrimination across clusters, the γ_{00} term represents the average difficulty of the item for reference group members across clusters, and the γ_{20} term represents the average uniform DIF across clusters. The τ_{00} term quantifies the variance in cluster-specific item difficulty and is similar to the item difficulty variance term introduced by Prowker and Camilli (2007). The τ_{11} term quantifies the variance in cluster-specific item discrimination. The parameter of most interest when studying DIF variance is τ_{22}, the variance of the cluster-specific uniform DIF coefficients. The cluster-specific deviations from the average DIF, the u_{2j} terms, represent the difference in performance between focal and reference groups in cluster j, relative to the average difference between focal and reference groups captured by γ_{20}. It is possible to constrain one or more of the terms in $\mathbf{T}$ to 0, which constrains the associated coefficients to be constant across groups. Constraining all elements of $\mathbf{T}$ to 0, for example, results in a model equivalent to the standard LR DIF model.

Conceptually, the RC-HLR model is similar to fitting separate LR DIF models within each cluster in a first stage of analysis, and then analyzing the estimated uniform DIF coefficients in a second stage of analysis (e.g., to estimate the variance among the coefficients). There are a number of potential advantages to using the RC-HLR model rather than the two-stage approach. First, in many cases the cluster sample sizes may be small, thus making estimation of separate β_2 statistics computationally difficult or very imprecise. By modeling the DIF coefficients as random variables rather than estimating them separately, the model is more parsimonious and leverages information from all clusters simultaneously. Subsequent analyses can use Empirical Bayes (EB) or related techniques to obtain predicted β_{2j} coefficients for each cluster that shrink imprecisely estimated DIF coefficients towards the overall average. Second, the HLR framework provides a natural way to extend the analyses by incorporating cluster-level covariates to study potential correlates or moderators of DIF across clusters. Chen and Zumbo (2017) used an ordinal HLR model to study country-level variation in DIF between male and female respondents on a reading attitude scale. Using a cumulative, ordinal HLR model, Chen and Zumbo found evidence that the magnitude of the gender DIF across countries was related to (i.e., moderated by) two indices measuring the level of development and gender inequality across countries. Although Chen and Zumbo included random effects in the model, the analysis and interpretation focused only on the average DIF coefficient estimates rather than identifying and explaining DIF variance.

As noted above, DIF variance could also arise if the DIF is mediated by an individual-level variable that differs systematically across clusters, or if the mediation varies across clusters. Cheng, Shao, and Lathrop (2016) describe the use of multiple-indicator, multiple-causes (MIMIC) models to study mediated DIF, with a focus on individual-level characteristics that may partially or completely mediate, and hence explain, an observed DIF effect between focal and reference groups. The RC-HLR model complements the mediated DIF framework by focusing on variation in the DIF effect across test administrations. Once DIF variance is identified, a mediated DIF MIMIC model could be used to understand the DIF, but as with standard LR DIF models, individual-level covariates could also be included directly in the RC-HLR model. Whether to focus on cluster-level covariates or individual-level covariates that can explain DIF variance would depend primarily on which research questions were of most interest to the researcher.

Random Intercept HLR Model and Multilevel DIF

French and Finch (2010) recommended using a random intercept HLR (RI-HLR) model in place of the standard LR DIF model when test-takers are clustered within groups. The RI-HLR is a constrained version of the RC-HLR model that constrains all elements of $\mathbf{T}$, except τ_{00}, to 0. Use of the RI-HLR model was motivated by a desire to improve the accuracy of estimated standard errors used to test for uniform DIF, under the assumption that the estimated standard errors in a LR DIF model would be too small if the clustering of test takers was not modeled appropriately, and lead to inflated Type I error rates. Using Monte Carlo simulations, French and Finch found that the LR and the RI-HLR models performed similarly when the grouping variable, g, varied within clusters; statistical power for the LR and RI-HLR models was nearly identical and Type I error rates remained at the nominal level for both models. When g varied between clusters (rather than within clusters), however, the LR DIF model had inflated Type I error rates that became higher as the sample size and intraclass correlation coefficient (ICC) of item responses increased. The RI-HLR model maintained Type I error rates at their nominal levels under these conditions. A subsequent study (Jin et al., 2014) found similar results, although with lower (but still inflated) Type I error rates for the LR model in conditions with a grouping variable that varied only between clusters. The difference in results was attributed to differences in the data generation methodology used in each simulation. Specifically, French and Finch manipulated the ICC of individual item responses, while Jin et al. manipulated the ICC of θ (the ability measured by the test), claiming that this was a more realistic representation of anticipated applications.

As described in Jin et al. (2014), the advantage of the RI-HLR DIF model relative to the LR DIF model is likely to depend primarily on ρ_y, the ICC of each item response, rather than ρ_θ, the ICC of ability. Here, $\rho_y = \tau_{00}/(\tau_{00} + \sigma^2)$ is the unconditional ICC of the studied item, where τ_{00} is the intercept variance for an HLR model with only an intercept (no x or g) and $\sigma^2 = \pi^2/3$, the variance of the standard logistic distribution (Goldstein, Brown, & Rasbash, 2002). The most relevant term impacting the LR DIF model is the conditional ICC of the studied item, $\rho_{y|x,g} = \tau^*_{00}/(\tau^*_{00} + \sigma^2)$, where τ^*_{00} is the intercept variance from the RI-HLR model including x and g. When x and g explain nearly all of the between-cluster variance in item responses (meaning $\rho_{y|x,g}$ is near 0), the standard LR and HLR models should perform similarly. As $\rho_{y|x,g}$ increases, the RI-HLR model should outperform the standard LR model. Although these prior studies found that the LR DIF model worked well when g was a within-cluster variable, neither study considered the case in which the magnitude of the within-cluster effect of g (DIF) varied across clusters. When there is DIF variance and the effect of g varies across clusters, this will tend to further increase $\rho_{y|x,g}$, potentially causing the standard LR DIF model to yield incorrect results. Adding variance to the effect of g causes the RI-HLR model to also be mis-specified and hence may cause the RI-HLR model to also yield incorrect results despite accounting for the clustering of test-takers.

This paper focuses on analyses with an observed, within-cluster g (i.e., an individual-level characteristic) because these are the types of group variables most frequently used in DIF analyses for large-scale assessments. When interest instead focuses on DIF due to a grouping variable that varies only between clusters, sometimes referred to as "cluster bias" (Jak, Oort, & Dolan, 2013), this would not result in the form of DIF variance described here. The papers by French and Finch (2010) and Jin et al. (2014) suggest that RI-HLR models provide one effective way to identify these DIF effects. See Jak, Oort, and Dolan (2014) or Jak et al. (2013) for a factor analytic approach to detecting cluster bias.

Alternative HLR DIF Models

Other applications of HLR models for DIF analysis (e.g., Cheong, 2006; Cheong & Kamata, 2013; De Boeck et al., 2011; Kamata, 2001; Swanson, Clauser, Case, Nungester, & Featherman, 2002) each differ in fundamental ways from the RC-HLR model described here. Using HLR, referred to as a hierarchical generalized linear model (HGLM), to fit an item response theory (IRT) model (Cheong, 2006; De Boeck et al., 2011) treats item responses as nested within test-takers and can be used to estimate the equivalent of a Rasch or 1-parameter logistic IRT model. This model has been used to study the effect of test context on DIF in prior analyses (Cheong, 2006), but treating DIF as a fixed rather than a random variable. The HGLM IRT model requires that a 1PL model fits the data and can face complexities in model identification when it is used to detect DIF (J.-H. Chen, Chen, & Shih, 2014; Cheong & Kamata, 2013). Swanson et al. (2002) proposed using HLR models to study variation in DIF across items within a test, rather than within items across clusters. In the model proposed by Swanson et al., level 1 represents individual examinees while level 2 represents items, allowing researchers to study the relationship between DIF magnitude and item features, while still assuming that each item has a single, stable DIF coefficient. The RC-HLR model described here can be used to study a distinct set of conceptual questions relative to these prior uses of HLR models.

Purpose of the Current Study

Use of the RC-HLR model to study DIF variance raises a number of practical challenges. Unlike linear multilevel models, closed-form solutions for calculating the likelihood are not available for HLR models, and estimation relies on approximations of the (log) likelihood (Bolker et al., 2009; Pinheiro & Chao, 2006). Testing the null hypothesis $H_0 : \tau_{22} = 0$ is complex, and commonly carried out using an approximate likelihood ratio χ^2 test (De Boeck et al., 2011). Use of the RC-HLR DIF model will likely encounter these issues in a number of ways. Although RC-HLR models should perform well with large sample sizes, researchers may be working with smaller sample sizes where quality of the estimates may depend on specific features of the data, such as the level of ICC among item responses or true ability. Finally, as with other LR DIF models (e.g., DeMars, 2010), the RC-HLR model is inherently mis-specified in the sense that x, the matching variable, contains measurement error.

Prior simulation studies have evaluated the performance of RI-HLR and RC-HLR models (e.g., Austin, 2010; Callens & Croux, 2005; Kim, Choi, & Emery, 2013; Moineddin, Matheson, & Glazier, 2007; Paccagnella, 2011; Schoeneberger, 2016), although not in the context of DIF. As Schoeneberger points out, differences in model specification, software, and estimation methods makes it difficult to draw generalizable conclusions across studies. This makes it difficult to predict how adequately the RC-HLR model will perform in conditions specific to a DIF analysis. In addition, none of these prior studies included conditions in which the covariates contained measurement error, nor did they appear to evaluate the performance of approximate χ^2 tests used to test the statistical significance of the variance components.

The purpose of this study was two-fold. First, using Monte Carlo simulations, to evaluate how accurately the RC-HLR model can detect uniform DIF and DIF variance across a range of conditions. The focus was on bias, Type I error rates, and statistical power when estimating the RC-HLR model using a widely available routine in the R software package (R Core Team, 2017). Second, using a real data analysis, to illustrate how the model can be applied and interpreted. The real data analysis also provides insights regarding plausible magnitudes of DIF variance researchers might expect to encounter in practice.

Simulation Study

A small Monte Carlo simulation was carried out to evaluate performance of the RC-HLR model when used to test for DIF in a single item of a 25-item test under a variety of conditions. The manipulated factors included sample size (number of clusters and number of test-takers per cluster), ICC of ability, mean difference in ability (i.e., "impact") between focal and reference groups, difficulty of the target item, and presence of uniform DIF and DIF variance in the target item. For each condition, a single target item in the test was systematically manipulated to have different combinations of difficulty, DIF, and DIF variance. Including conditions where the target item did not have true DIF or DIF variance allowed Type I error rates to be studied in addition to statistical power. All factors were included in a fully crossed design resulting in a total of 4 (sample size) × 2 (ICC) × 2 (impact) × 2 (DIF) × 2 (DIF Variance) × 3 (item difficulty) = 192 conditions. All data generation and analyses were conducted in R (R Core Team, 2017), relying in part on the SimDesign package (Chalmers, 2018).[2]

Simulation Conditions

Sample Size. The number of clusters was set at either $J = 25$ or $J = 100$ and sample size per cluster was set at either $N = 10$ or $N = 40$. These values were crossed resulting in 4 conditions that represent a wide range of total sample sizes from 250 to 4,000.

There were an equal number of focal and reference group students in each cluster for all conditions.

Intraclass Correlation Coefficient (ICC) of Ability. To study the impact of heterogeneity of ability across groups, the ICC of true student ability across clusters was set at either $\rho_\theta = 0.05$ or $\rho_\theta = 0.30$. These represent a range from low to high based on ICC values estimated with real test score data (e.g., Hedges & Hedberg, 2007).

Impact. To simulate differences in average ability between reference and focal groups, impact was set at either $\delta = 0$ or $\delta = 1$, corresponding to either no mean difference or a large mean difference of 1 standard deviation favoring the reference group.

Target Item DIF. The target DIF item was simulated either to have no uniform DIF or an average uniform difference in item difficulty of 0.6 in the logit metric, so that the true value of γ_{20} was either 0.0 or -0.6. The condition with 0 average uniform DIF allows Type I error rates to be evaluated, and the condition with non-zero DIF allows the relative power and bias of the methods to be compared. The value 0.6 was selected to be comparable to the values used in prior HLR and LR DIF studies (e.g., French & Finch, 2010; French & Maller, 2007; Hidalgo et al., 2014; Jin et al., 2014; Jodoin & Gierl, 2001). The target DIF item had a discrimination of $a = 1$ in both groups. A DIF magnitude of $b_{focal} - b_{reference} = 0.6$ in a 2PL-IRT model corresponds to an area between the item response curves of 0.6, using Raju's (1988) formula. Converting this to the Educational Testing Service (ETS) Δ scale yields $\Delta = 2.35 * -0.6 = -1.41$, which would be classified as a B DIF item (moderate DIF), assuming the DIF is significantly different from 0 (Roussos, Schnipke, & Pashley, 1999).

Target Item DIF Variance. The target DIF item was also manipulated to either have constant DIF across clusters or non-zero DIF variance across clusters. The true value of τ_{22} was set to either 0.0 or 0.80. Because prior studies do not exist for comparison of this parameter, the value was chosen based on results in the real data analysis.

Target Item Difficulty. Prior studies have found that item difficulty and discrimination parameters can affect the efficacy of HLR or LR DIF detection (Jin et al., 2014; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993). The target DIF item was set to have a difficulty in the reference group of either $b_r = [-0.75, 0.00, 0.75]$.

Additional Item Parameters. Item parameters for the remaining 24 non-target DIF items were held constant and DIF-free across all conditions. These item parameters were randomly sampled from a set of operational item parameters from a statewide 8th grade mathematics test (Education, 2016). The items were originally scaled with a 3PL model, but only the a and b parameters were used in this study. The randomly sampled parameters were re-scaled to have an average difficulty of $b = 0$ and average discrimination of $a = 1$.

Data Generation

For each condition defined by a combination of the above factors, 200 replications of data were simulated. In each replication, true ability values θ_{ij} were generated for each simulated examinee i in cluster j. To induce an ICC of ρ_θ, random normal deviates $e_{ij} \sim N\left(0, (1 - \rho_\theta)\right)$ were simulated for each examinee and combined with normally distributed cluster random effects $\mu_j \sim N\left(0, \rho_\theta\right)$, so that the distribution of θ_{ij} was standardized with a marginal mean and variance of 0 and 1. To simulate non-zero group impact of δ, the e_{ij} values were simulated with mean $\delta/2$ for the reference group and $-\delta/2$ for the focal group, and the variance of e_{ij} was set equal to $1 - \rho_\theta - (p_{ref} * (1 - p_{ref}) * \delta^2)$, where p_{ref} is the proportion of students in the reference group (fixed to 0.5 in all conditions), to maintain the population mean and variance of θ_{ij} at 0 and 1. This simulation procedure amounts to randomly sampling N_j students from each of J randomly sampled clusters in each replication.

Item responses were simulated based on a 2PL-IRT model (Embretson & Reise, 2000). The probability that examinee i in cluster j answers item k correctly, denoted p_{ijk}, conditional on θ_{ij}, item discrimination a_k, item difficulty b_k, and cluster-specific DIF $(\gamma_{20k} + u_{2jk})$, and group membership indicator g_{ij} ($g_{ij} = 1$ for focal group and 0 for reference group) is

$$p_{ijk} = \frac{\exp\left(a_k * (\theta_{ij} - b_k - g_{ij} * [\gamma_{20k} + u_{2jk}])\right)}{1 + \exp\left(a_k * (\theta_{ij} - b_k - g_{ij} * [\gamma_{20k} + u_{2jk}])\right)}. \tag{6}$$

Item responses y_{ijk} for each examinee and each item were simulated by generating an independent uniform variable z_{ijk} on the interval $[0, 1]$ and setting

$$y_{ijk} = \begin{cases} 1, & \text{if } z_{ijk} < p_{ijk} \\ 0, & \text{if } z_{ijk} \geq p_{ijk} \end{cases}. \tag{7}$$

Model Specificaion and Estimation

For each replication of each condition, the target DIF item was tested for significant average uniform DIF and DIF variance. Three models were used to test for average DIF: the standard LR DIF model, the RI-HLR model with no random coefficients, and the RC-HLR model with τ_{00} and τ_{22} freely estimated and all other elements of $\mathbf{T}$ constrained to 0. A Wald statistic based on the estimated standard error was used to test the null hypothesis of no DIF for each of these three models. The LR DIF models were estimated using the `glm()` function in R while the RI-HLR and RC-HLR models were estimated using the `glmer()` function in the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015), using the Laplace approximation to the likelihood. The standardized total score across all items, centered at the grand mean, was used as the matching variable in all models. The group indicator was set to 0 for members of the reference group and 1 for members of the focal group.

The RC-HLR model was also used to test for DIF variance in the target item. To test the null hypothesis that DIF variance was 0, $H_0 : \tau_{22} = 0$, an approximate likelihood ratio test was used. This test was constructed by fitting both the RC-HLR model and the reduced RI-HLR model that constrains $\tau_{22} = 0$. Letting LL_0 be the approximate log-likelihood of the reduced model and LL_1 be the approximate log-likelihood of the full model, an asymptotically χ^2 statistic can be constructed as $\Gamma = -2 \ln (LL_0/LL_1)$. This statistic can be compared to a 50:50 mixture of a χ^2 distribution with 0 and 1 degrees of freedom to adjust for the fact that the null hypothesis value of 0 for τ_{22} is on the boundary of the parameter space, because it is required that $\tau_{22} \geq 0$ (De Boeck et al., 2011).

If either the LR, RI-HLR, or RC-HLR model failed to converge in a given replication, a new sample of data was generated.[3] This was repeated as necessary to achieve 200 replications for which all three models converged in each condition so that all models could be compared in each condition. The number of attempted replications needed in order to obtain 200 replications with convergence of all three models was recorded to indicate conditions for which convergence may be problematic in practice. The nominal Type I error rate was set at $\alpha = 0.05$ for all tests. If the true proportion of Type I errors is $p = 0.05$, then observed Type I error rates with a sample size of $n = 200$ should be within $\pm 2\sqrt{p(1-p)/n} \approx 0.03$ approximately 95% of the time. Type I error rates were considered acceptable if the observed Type I error rate remained at or below 0.08. This is also similar to the recommendation of Bradley (1978) that a statistical test be considered robust if it maintains Type I error rates within $\pm 0.5 * \alpha$, where α is the nominal α-level.

Results

There were two primary sets of outcomes of interest: accuracy of hypothesis tests and point estimates for the average uniform DIF coefficient, γ_{20}, across models, and the accuracy of hypothesis tests and point estimates for the DIF variance term, τ_{22}, using the RC-HLR model. Convergence results and patterns of item-level ICC's (ρ_y and $\rho_{y|x,g}$) were also recorded.

Convergence and Observed ICC's. The standard LR model converged in all replications. The RI-HLR model failed to converge in only 0.11 percent of replications. The RC-HLR model failed to converge in 9.21 percent of replications. Achieving 200 replications for which all three models converged required generating anywhere from 201 to 296 (and an average of 218.42) total replications across conditions, with higher rates of non-convergence for the RC-HLR model in conditions in which the target item was easy, had no true average uniform DIF, and had non-zero DIF variance. With no DIF variance, the average unconditional ICC's (ρ_y) of the target item were 0.010 and 0.055 when the ICC of ability was 0.05 and 0.30, respectively; with non-zero DIF variance, the average unconditional ICC's were 0.036 and 0.082, respectively. The average conditional ICC's ($\rho_{y|x,g}$, estimated from the RI-HLR model) were 0.007 and 0.044 for conditions without and with DIF variance, respectively. This suggests that unconditional item ICC's were affected by both DIF variance and the ICC of ability, while the conditional ICC was affected primarily by the DIF variance.

Type I Error Rate of Average Uniform DIF Test. Figure 1 displays Type I error rates for the test of average uniform DIF across all models and conditions in which the true average uniform DIF was 0. Within each panel of Figure 1, each shaded bar represents the observed Type I error rate for a particular model and sample size combination. Each bar represents 200 replications and conditions in which Type I error rates were greater than 0.08 are indicated with a "*" printed above the bar. Type I error rates were primarily affected by the presence of DIF variance and sample size. When there was no DIF variance, all three models maintained accurate Type I error rates with only one exception (with J=25 clusters of size N=10, there was one condition for which Type I error rates for the LR and RI-HLR models were slightly inflated at 0.09).

When the studied item included DIF variance, Type I error rates could become substantially inflated for the LR and RI-HLR models. Both models had similar Type I error rates. Table 1 summarizes observed Type I error rates by DIF variance condition, sample size, and model. Type I error rates were above the 0.08 cutoff in 39 out of the 48 conditions for both the LR and RI-HLR models, and were higher in conditions with larger sample sizes. In the largest sample size condition, Type I error rates were as high as 0.380 for the LR model and 0.365 for the RI-HLR model. Type I error rates for the RC-HLR model were inflated for a smaller number of conditions (7 out of the 48 conditions), but the average Type I error rates were much lower; the highest Type I error rate for the RC-HLR model was 0.090. In the largest sample size condition, the average observed Type I error rate of the RC-HLR model was 0.056, while it was 0.260 and 0.265 for the LR and RI-HLR models, respectively. The ICC of ability and mean difference of ability did not have large effects on the Type I error rates, while the difficulty of the target item had a small effect (Type I error rates were slightly higher when the target item was very easy or very difficult).

To further understand the inflated Type I error rates, the standard error ratio in each condition was computed as the average estimated standard error divided by the observed standard deviation of the estimated DIF coefficient across all 200 replications. When there was no DIF variance, the estimated standard errors were similar in magnitude to the observed empirical standard errors for all models and all conditions. The average ratios for the LR, RI-HLR and RC-HLR models were 0.998, 0.997, and 1.009, respectively, ranging from a minimum of 0.906 (for the LR model) to 1.130 (for the RC-HLR model). When there was DIF variance in the target item, the estimated standard errors remained similar to the empirical standard errors for the RC-HLR model, although they were slightly too small. The average standard error ratio across all conditions with DIF variance for the RC-HLR model was 0.977 (Min=0.872, Max=1.130). For the LR and RI-HLR models, the standard errors were substantially underestimated, particularly with larger sample sizes. The average standard error ratio for the LR model was 0.797 (Min=0.621, Max=0.984) and for the RI-HLR model was 0.791 (Min=0.613, Max=0.966). This indicates substantial underestimation of the sampling error in the average uniform DIF coefficient, a widely recognized problem in models that do not adequately account for clustering in multilevel data (e.g.; Hox, 2010).

Power of Average Uniform DIF Test. The test of average uniform DIF was considered accurate and sufficiently powered if: a) Type I error rates for the same condition (without average DIF) were not inflated, and b) observed power achieved a level of 0.80 or greater. Table 2 summarizes the mean, median, minimum, and maximum statistical power by DIF variance condition and model. The values N0, N1, and N2 represent, respectively, the total number of conditions summarized in each row, the number of conditions with observed power greater than 0.80, and the number of conditions with observed power greater than 0.80 and Type I error rates less than or equal to 0.08. With no DIF variance in the target item, all models performed similarly and statistical power was above 0.80 for all but the smallest sample size condition (J=25 and N=10), for which power reached a minimum of 0.405 for the RC-HLR model and 0.425 for the LR and RI-HLR models.

Figure 2 shows the statistical power for the test of average uniform DIF for all conditions with true DIF variance. As in Figure 1, conditions for which Type I error rates were inflated above 0.08 are indicated with a "*" printed above the bar and should be interpreted cautiously. Figure 2 shows that statistical power was similar for all three models and varied primarily as a function of sample size. Ignoring incorrect Type I error rates, Table 2 shows that power was slightly lower for all three models relative to conditions with no DIF variance, and average power for the RC-HLR model was slightly lower than power for the LR and RI-HLR models. Although the LR and RI-HLR models had observed statistical power greater than 0.80 in 30 of the 48 conditions with DIF variance, relative to 25 such conditions for the RC-HLR model, the LR and RI-HLR model only correctly controlled Type I error rates in three of these conditions. The RC-HLR model, however, maintained correct Type I error rates and had high power in 22 of the 48 conditions. For the two conditions with equal total sample size of 1000, power for the RC-HLR model tended to be greater with in the conditions with more clusters (J=100 and N=10) rather than more students per cluster (J=25 and N=40).

Bias of Average DIF Estimate. Figure 3 summarizes the bias in estimated DIF coefficients across DIF conditions, target item difficulty, and model. Each boxplot summarizes the bias in estimated DIF for 16 different conditions for a single model. Each column presents results for a different model, while the rows indicate different combinations of average DIF and DIF variance. All models yielded nearly unbiased estimates of average uniform DIF when there was no DIF variance. When there was DIF variance, the RC-HLR model continued to produce nearly unbiased estimates of average uniform DIF, while there was systematic bias in the average uniform DIF estimates produced by the LR and RI-HLR models. The bias for the LR and RI-HLR models varied primarily as a function of the target item difficulty and average uniform DIF. Due to the coding of focal and reference groups, positive bias indicates overestimating the relative advantage for the focal group while negative bias indicates overestimating the relative advantage for the reference group. When there was no true average uniform DIF, the LR and RI-HLR models had negative bias when the target item was easy and positive bias when the target item was difficult. When true average DIF was present, the relative pattern of the bias was similar, but shifted upwards, with almost no bias when the target item was easy, and larger positive bias when the target item was difficult. Because the true average DIF was -0.6 in these conditions, the LR and RI-HLR models tended to underestimate the magnitude of the DIF favoring the reference group.

Type I Error Rate, Power, and Bias of DIF Variance. Figure 4 displays significance rates (Panel A) and bias (Panel B) of the estimated DIF variance parameter τ_{22}. The left portion of Panel A displays Type I error rates while the right portion displays statistical power. Because results varied primarily as a function of sample size, each boxplot summarizes the results for 48 conditions of the indicated sample size and DIF variance condition. Type I error rates (displayed in the left portion of Panel A) were always maintained at or very close to the nominal level of 0.05 (indicated by the horizontal dashed line); the highest observed Type I error rate was 0.055. The test of DIF variance tended to be conservative, with an overall average Type I error rate of 0.017. Statistical power (shown in the right portion of Panel A) was low in the smallest sample size conditions (Mean=0.305, Min=0.215, Max=0.375), but generally above 0.80 for all other sample size conditions. Unlike the test for average DIF, for equal total sample sizes of 1000, power was higher in conditions with larger within-cluster sample sizes (J=25, N=40) than for conditions with a larger number of clusters (J=100, N=10).

Regarding accuracy of the DIF variance estimates, the left portion of Panel B shows there was positive bias in the estimated DIF variance for conditions in which the true DIF variance was 0. This was not an unexpected result, and occurs because the estimated DIF variance must be greater than or equal to 0. The positive bias decreased as sample size increased. The right side of Panel B shows bias when the true DIF variance was 0.80, and indicates there was negative bias in the estimated DIF variance that decreased as sample size increased. In the smallest sample size condition, the average bias was -0.133, meaning the DIF variance was underestimated by approximately 16.6%. As with statistical power, for equal total sample sizes of 1000, more accurate estimates of the DIF variance were obtained in conditions that had larger within-cluster sample sizes rather than larger numbers of clusters. Decreasing bias as within-cluster sample size increases is consistent with with prior results for variance components estimated in HLR models using the Laplace approximation (e.g., Joe, 2008). In the largest sample size conditions, average bias was -0.038, or approximately 4.7% of the true value.

Real Data Example

Data and Methods

To demonstrate application of the RC-HLR DIF model to real data, a DIF analysis was conducted using data from a high school mathematics test. The data are based on responses to a 27-item mathematics test form administered to a national sample of 10th graders as part of the Education Longitudinal Study of 2002 (Ingels et al., 2004). All items were scored dichotomously as correct/incorrect. The sample of all students completing the test was reduced to students who had non-missing data for student gender (male or female) and school identifiers, completed the full in-person questionnaires and achievement tests on the scheduled testing date, answered at least half of the questions on the test form, and had at least 10 students at their school complete the test form. Due to these sample restrictions, this sample is not necessarily representative of the national 10th grade population of students. Data were obtained from a restricted-use data license provided by the National Center for Education Statistics (NCES), and all sample sizes are rounded to the nearest 10 to comply with NCES reporting requirements.

The final sample included data for 850 students (470 male and 380 female) across 70 schools. Male students answered slightly more questions correctly (Mean=12.5, SD=4.5) than female students (Mean=11.4, SD=4.0). The ICC of observed total scores across schools was approximately 0.126. The average unconditional ICC of the items responses was $\bar{\rho}_y = 0.077$ and the average conditional ICC was $\bar{\rho}_{y|x,g} = 0.066$ (see Table 3 for all values). These values are similar to those observed in the simulated item responses with DIF variance above and those reported by Jin et al. (2014) when generating data from a 2PL IRT model with a non-zero ICC of true ability. This suggests there may be some benefit to the use of HLR DIF models over LR DIF models for these data.

Average uniform DIF between male and female students was tested for each item using the same procedures for the LR, RI-HLR, and RC-HLR models as described above in the simulations. DIF variance was estimated with the RC-HLR model. The matching variable was the observed total score (centered at the grand mean and scaled to have a standard deviation of 1) and the grouping variable was a binary indicator for student gender equal to 1 if a student was female and 0 if a student was male. The hypothesis tests were carried out with a nominal $\alpha = 0.05$.

Results

Table 3 displays the estimated DIF coefficients for the LR, RI-HLR and RC-HLR models, the estimated DIF variance based on the RC-HLR model, the unconditional and conditional ICC for each item, and the overall proportion correct for each item. Significant average uniform DIF coefficients and DIF variance estimates are indicated with a "*". Results are sorted based on the magnitude of the estimated LR uniform DIF coefficient to facilitate interpretation. Negative DIF estimates indicate items that favor male students, while positive estimates indicate items that favor female students. The RC-HLR model did not converge for 4 items. For three of the items (items 3, 15, and 22) the estimated DIF variance at the final iteration was very close to 0, and the results for these items are based only on the LR and RI-HLR model that constrains DIF variance to 0 (i.e., these items are assumed to have no DIF variance). The fourth item (item 10) was re-estimated using an RC-HLR model that allowed a non-zero covariance between random intercepts and random DIF coefficients.

Using the LR model, 10 of the 27 items would be flagged as potentially having DIF. Using the RI-HLR and RC-HLR models, seven and five items were flagged, respectively, and these were subsets of the original 10 flagged items. The direction of DIF was the same across all three models, and the absolute magnitude of the DIF estimates tended to be similar or slightly larger for the RC-HLR model. Of the five items flagged for DIF by the RC-HLR model, two were also flagged as having significant DIF variance. Four additional items not flagged for average DIF by the RC-HLR model (including one not flagged for average DIF by any model) were also flagged for significant DIF variance. The items flagged for DIF variance had a range of proportion correct values from 0.244 (item 11) to 0.789 (item 1). With a nominal $\alpha = 0.05$, one would expect to flag only 1-2 items (out of 27) for DIF or DIF variance by chance.

These results suggest there may be heterogeneity in the observed DIF for six of the items in this test. Specifically, half of the 10 items that would be flagged by a standard LR DIF analysis as having uniform DIF, and one additional item not flagged by the LR model, displayed evidence of DIF variance that requires further consideration. Item 11, for example, had an estimated DIF coefficient of -1.084 (using the RC-HLR estimate), but the estimated standard deviation of DIF across schools was $0.977 = \sqrt{0.955}$. If the DIF coefficients are normally distributed (an assumption of the RC-HLR model), then approximately 15% of schools would be expected to have uniform DIF one standard deviation or more below the estimated value of -1.084, corresponding to a DIF coefficient of -2.061, a very large DIF estimate. On the other hand, approximately 15% of schools would be expected to have DIF coefficients one standard deviation or more above the mean value, corresponding to -0.107, a negligible amount of DIF. The logistic regression DIF coefficient can be converted to the ETS Δ scale by calculating $\hat{\Delta} = 2.35 * \hat{\beta}_2$ (Monahan, McHorney, Stump, & Perkins, 2007); note this expression is the negative of that used by Monahan et al. due to a difference in coding the focal and reference groups. The results suggest that across different schools this item could have either small (category A) or large (category C) DIF. These represent extremes from very unequal to not substantially different performance, and indicate that summarizing the item DIF with a single average statistic may not adequately capture the patterns in item responding.

These results also display patterns consistent with the simulation results. The greater number of items flagged for DIF with the LR and RI-HLR models is consistent with the inflated Type I error rates and these may represent false positives. The items flagged for both average DIF and DIF variance by the RC-HLR model (items 1 and 11) had larger absolute average uniform DIF estimates from the RC-HLR model than from the LR or RI-HLR models. Finally, the Pearson correlations between estimated DIF variance (excluding items for which DIF variance was constrained to be 0) and ρ_y and $\rho_{y|x,g}$ across items were 0.612 and 0.622, respectively. This further suggests that the magnitude of ρ_y and $\rho_{y|x,g}$ may be good initial indicators of DIF variance.

In practice researchers would need to decide which set of model estimates to interpret. Based on the simulation results and empirical analyses, a two-step modeling approach could be used. First, the full RC-HLR model would be used to test for significant DIF variance. If there is evidence of DIF variance, then results from the RC-HLR model should be used. If there is no evidence of DIF variance, then researchers must select between either the RI-HLR or LR DIF models. The simulations above and prior studies (French & Finch, 2010; Jin et al., 2014) suggest that when there is no DIF variance and the DIF grouping variable is a within-cluster variable, both the RI-HLR and LR DIF models should yield accurate parameter estimates and hypothesis tests. The results in Table 3 indicate that the LR model flags two items (21 and 14) for DIF that the RI-HLR model would not flag and that do not show DIF variance. This result is consistent with the concern that the LR model may underestimate the standard errors of the DIF coefficient when there is a multilevel structure in the data. Hence the RI-HLR model is likely the more appropriate (and also more conservative) model to interpret when there is no evidence of DIF variance, but there is a relevant multilevel data structure. As discussed below, future research should also investigate potential effect size criteria that could be included to complement the significance tests used to identify and interpret DIF.

Discussion

This paper described how RC-HLR models can be used to quantify and test for DIF variance, an indicator that item DIF varies across test administrations. A simulation study was used to evaluate how well the RC-HLR model works for this purpose across a range of conditions, and an empirical example was used to demonstrate the use and interpretation of HLR models for this purpose. The simulation results suggest the RC-HLR model is a promising approach to further our understanding of DIF and the empirical analyses found evidence of variance in the DIF between male and female students across schools, indicating a potential avenue for future investigation.

The simulations documented that when the average uniform DIF was 0 but there was nonzero DIF within some clusters, Type I error rates for the standard LR or RI-HLR models could become substantially inflated. The inflated Type I error rates tended to increase with sample size, particularly as the within-cluster sample size increased. The RC-HLR model substantially reduced these false positive rates and generally maintained Type I error rates at their nominal levels. For the DIF conditions studied here, the RC-HLR model had statistical power that was only slightly lower than the power observed for the LR and RI-HLR models. These results suggest that when there is DIF variance, the RC-HLR model trades a slight reduction in statistical power for a substantial increase in the number of conditions with both adequate power and correctly controlled Type I error rates. When there was no DIF variance, all three models had very similar Type I error and statistical power rates.

The inflated Type I error rates in this study are difficult to compare to prior DIF simulations, which did not include DIF variance. The within-cluster group condition used by French and Finch (2010) and Jin et al. (2014) were similar to the conditions in this simulation without DIF variance, and hence the finding of correct Type I error rates across all models replicates earlier studies. The lack of effect of the ICC of ability on Type I error rates was likely because after conditioning on x and g, there is little remaining between-cluster variation in item responses (e.g., Jin et al., 2014). The conditions with DIF variance differ fundamentally from both the within-cluster and between-cluster conditions used in earlier studies, in which the DIF was modeled as a constant effect of g. When the magnitude of the DIF varies across clusters, conditioning on a fixed g alone cannot correctly account for the structure of the data, and as a result even the RI-HLR models can have inflated Type I error rates. Finally, although it was anticipated that mean differences in ability might affect Type I error rates, this was not found. The lack of effect of mean differences in ability is consistent with prior studies in which there was only a single DIF item on a test (e.g., Hidalgo et al., 2014).

One could argue that in cases with zero average uniform DIF but non-zero DIF variance, flagging an item as having DIF is not a false positive, because in some administrations (clusters) the magnitude of the DIF is non-zero. In these cases, however, researchers would falsely conclude from using an LR or RI-HLR model that the overall effect of the DIF was unidirectional and would fail to identify the heterogeneity in the DIF across administrations. If an item has characteristics that lead to heterogeneity in DIF across administrations, but researchers are searching for an item characteristic that could explain a consistent unidirectional DIF magnitude, this could be an additional reason for the common occurrence of finding significant DIF that has no clear cause or explanation (Angoff, 1993).

In addition to incorrectly flagging items as having non-zero average uniform DIF, failing to model the DIF variance can also lead point estimates of DIF to be inaccurate. In the simulation, DIF favoring the reference group was underestimated for difficult items and overestimated for easier items when there was DIF variance. To gain intuition into this result, consider a very difficult test item. If there is true DIF variance in this item across administrations, the proportion of focal group students answering the item correctly in administrations with positive DIF (favoring the focal group) will tend to be relatively higher, and in some cases this increase could be substantial. In administrations with negative DIF (favoring the reference group) the proportion of focal group students answering the item correctly will decrease, but if the item is already very difficult the magnitude of the decrease is bounded within a smaller range. On average, difficult items will tend to look relatively easier for the focal group, while the opposite phenomenon can occur for easier items.

Finally, the simulations documented that the approximate likelihood ratio test used for DIF variance had adequate power for moderate to large sample sizes and should yield point estimates of the variance accurate enough to support further study. As expected, estimates of DIF variance were slightly positively biased when the true DIF variance was 0, but estimates were flagged as statistically significant at or below the nominal α level. Although there was negative bias in estimated variances when the true variance was non-zero, the relative magnitude of the bias was small and decreased as sample size increased (particularly within-cluster sample size), consistent with prior simulation studies evaluating HLR models estimated using a Laplace approximation (e.g., Austin, 2010; Joe, 2008; Paccagnella, 2011; Schoeneberger, 2016).

The empirical data analysis provided evidence of DIF variance in the studied items and indicated that the LR and RI-HLR models may be incorrectly flagging items for average DIF or inaccurately estimating the magnitude of the DIF. An important next step would entail following up the DIF analyses to better understand whether there are relevant features of the school contexts, such as instructional practices or features of the test-taking session, that can explain the variation in DIF coefficients. There are a few approaches that could be used for such analyses. First, one could systematically search for common characteristics among the items displaying significant DIF variance to better understand which types of items are displaying DIF variance. Second, the RC-HLR model could be extended to include cluster-level covariates to determine whether these covariates are correlated with the magnitude of the DIF coefficient in each cluster. This latter option represents a more confirmatory approach to studying DIF variance. While these analyses would not necessarily allow for causal inferences about DIF, they could provide suggestive evidence for potential cluster-level moderators of DIF. The RC-HLR model, and the LR framework for DIF more generally, also allow one to include additional covariates at the cluster or individual level, and coould potentially be used to explore DIF mediators as well (Cheng et al., 2016).

Limitations

As with any simulation study there are important limitations and next steps to highlight. These include model estimation techniques, use of effect sizes for flagging items, generalizability of the simulation conditions, and extensions of the RC-HLR framework. In the simulation, the covariance term between the random intercept and random DIF coefficient were constrained to 0 (to match the data generating procedure), but convergence problems remained for some conditions and in the real data analysis. It thus may be worth evaluating alternative model specifications and estimation algorithms. This could include comparing the results to other software packages or to the use of Bayesian estimation frameworks that incorporate prior distributions, and which may overcome some of the small sample problems encountered.

Incorporating effect sizes into the criteria for flagging DIF items would be a useful extension that could potentially mitigate inflated false positive results and avoid overreliance on significance tests. This study did not use an effect size criteria because there is little consensus about the optimal effect size criteria to use even in standard LR DIF models (e.g., French & Maller, 2007; Hidalgo et al., 2014; Jodoin & Gierl, 2001; Zumbo, 1999). Use of an effect size criteria for flagging practically significant DIF variance will require further study regarding the typical amount of DIF variance found in practice. The use of a purification strategy to purify the matching score prior to flagging items is another factor that could be investigated. Although the simulation conditions were carefully selected to include scenarios that could be commonly encountered (and were similar to the real data conditions), evaluation of additional factors, including a wider range of sample sizes, DIF variance magnitudes, uniform DIF magnitudes, and proportion of DIF items would help to inform future use of the RC-HLR model. These additional conditions were not included in this study in order to maintain focus and a manageable number of results. Finally, extensions to the RC-HLR model could be explored, including the use of random coefficients on the discrimination paramters and the inclusion of nonuniform DIF coefficients.

The real data analysis provided an example of the RC-HLR applied in practice and suggested that there may be scenarios in which DIF variance is a relevant concern. Future analyses could focus both on evaluating whether there is evidence of DIF variance in additional settings and on determining which (if any) cluster-level variables may be able to explain this DIF variance. The correlation between DIF variance and the item ICC's suggests that estimating the unconditional and conditional ICC's may be a good initial indicator of DIF variance that researchers can use when planning a DIF analysis.

Conclusion

The proposed application of RC-HLR models to study DIF variance is intended to provide researchers with a new method for quantifying and understanding heterogeneity in item responding across test administrations. The RC-HLR model can potentially improve test fairness by more accurately identifying and quantifying uniform DIF and by providing a method for quantifying DIF heterogeneity. By directly modeling both the manifest DIF grouping variable and the observed clustering of test-takers within test administrations, the RC-HLR model provides a complement to prior HLR DIF models and to other methods for studying DIF heterogeneity, either within manifest DIF groups (Cheng et al., 2016; e.g., Oliveri, Ercikan, & Zumbo, 2014) or across unobserved strata (i.e., latent classes; Oliveri, Ercikan, Zumbo, & Lawless, 2014; Zumbo et al., 2015). The RC-HLR model provides an additional tool that researchers can use to articulate and evaluate a wider range of test score interpretations, with a particular focus on identifying cases where features of the testing situation might moderate DIF as conceptualized by the ecological model of item responding (Zumbo et al., 2015).

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing.* Washington, D.C.: American Educational Research Association.

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, N.J.: John Wiley & Sons.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Anrig, G. R. (1987). ETS on "Golden Rule". *Educational Measurement: Issues and Practice*, *6*(3), 24–27. doi:10.1111/j.1745-3992.1987.tb00503.x

Austin, P. C. (2010). Estimating multilevel logistic regression models when the number of clusters is low: A comparison of different statistical software procedures. *The International Journal of Biostatistics*, *6*(1). doi:10.2202/1557-4679.1195

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). doi:10.18637/jss.v067.i01

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135. doi:10.1016/j.tree.2008.10.008

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. doi:10.1111/j.2044-8317.1978.tb00581.x

Callens, M., & Croux, C. (2005). Performance of likelihood-based estimation methods for multilevel binary regression models. *Journal of Statistical Computation and*

Simulation, *75*(12), 1003–1017. doi:10.1080/00949650412331321070

Chalmers, P. (2018). *SimDesign: Structure for organizing monte carlo simulation designs.* Retrieved from https://CRAN.R-project.org/package=SimDesign

Chen, J.-H., Chen, C.-T., & Shih, C.-L. (2014). Improving the control of type I error rate in assessing differential item functioning for hierarchical generalized linear model when impact is presented. *Applied Psychological Measurement*, *38*(1), 18–36. doi:10.1177/0146621613488643

Chen, M. Y., & Zumbo, B. D. (2017). Ecological framework of item responding as validity evidence: An application of multilevel DIF modeling using PISA data. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (Vol. 69, pp. 53–68). Cham: Springer. doi:10.1007/978-3-319-56129-5_4

Cheng, Y., Shao, C., & Lathrop, Q. N. (2016). The mediated MIMIC model for understanding the underlying mechanism of DIF. *Educational and Psychological Measurement*, *76*(1), 43–63. doi:10.1177/0013164415576187

Cheong, Y. F. (2006). Analysis of school context effects on differential item functioning using hierarchical generalized linear models. *International Journal of Testing*, *6*(1), 57–79. doi:10.1207/s15327574ijt0601_4

Cheong, Y. F., & Kamata, A. (2013). Centering, scale indeterminacy, and differential item functioning detection in hierarchical generalized linear and generalized linear mixed models. *Applied Measurement in Education*, *26*(4), 233–252. doi:10.1080/08957347.2013.824453

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., … others. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*(12), 1–28. Retrieved from http://ppw.kuleuven.be/okp/_pdf/DeBoeck2011TEOIR.pdf

DeMars, C. E. (2010). Type I error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement*, *70*(6), 961–972. doi:10.1177/0013164410366691

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*(4), 355–368. doi:10.1111/j.1745-3984.1986.tb00255.x

Education, M. D. of. (2016). *2015 MCAS and MCAS-Alt technical report.* Retrieved from http://www.mcasservicecenter.com/documents/MA/Technical%20Report/TechReport_2015.htm

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement*, *47*(3), 299–317. doi:10.1111/j.1745-3984.2010.00115.x

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement*, *67*(3), 373–393. doi:10.1177/0013164406294781

Goldstein, H., Brown, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics*, *1*(4), 223–231.

doi:http://dx.doi.org/10.1207/S15328031US0104_02

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park: SAGE Publications.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. doi:10.3102/0162373707299706

Hidalgo, M. D., Gomez-Benito, J., & Zumbo, B. D. (2014). Binary logistic regression analysis for detecting differential item functioning: Effectiveness of R2 and delta log odds ratio effect size measures. *Educational and Psychological Measurement*, *74*(6), 927–949. doi:10.1177/0013164414523618

Holland, P. W., & 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 Associates.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, N.Y.: Routledge.

Ingels, S. J., Pratt, D. J., Rogers, J. E., Siegel, P. H., Stutts, E. S., & Owings, J. A. (2004). *Education longitudinal study of 2002: Base year data file user's manual*. Washington, D.C.: US Department of Education National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubs2004/2004405.pdf

Inzlicht, M., & Ben-Zeev, T. (2000). A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males.

Psychological Science, *11*(5), 365–371. doi:10.1111/1467-9280.00272

Inzlicht, M., & Ben-Zeev, T. (2003). Do high-achieving female students underperform in private? The implications of threatening environments on intellectual processing. *Journal of Educational Psychology*, *95*(4), 796–805. doi:10.1037/0022-0663.95.4.796

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(2), 265–282. doi:10.1080/10705511.2013.769392

Jak, S., Oort, F. J., & Dolan, C. V. (2014). Measurement bias in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(1), 31–39. doi:10.1080/10705511.2014.856694

Jin, Y., Myers, N. D., & Ahn, S. (2014). Complex versus simple modeling for DIF detection: When the Intraclass Correlation Coefficient (rho) of the studied item is less than the rho of the total score. *Educational and Psychological Measurement*, *74*(1), 163–190. doi:10.1177/0013164413497572

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329–349. doi:10.1207/S15324818AME1404_2

Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. *Computational Statistics & Data Analysis*, *52*(12), 5066–5074. doi:10.1016/j.csda.2008.05.002

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of*

Educational Measurement, *38*(1), 79–93. doi:10.1111/j.1745-3984.2001.tb01117.x

Kim, Y., Choi, Y.-K., & Emery, S. (2013). Logistic regression with multiple random effects: A simulation study of estimation methods and statistical packages. *The American Statistician*, *67*(3), 171–182. doi:10.1080/00031305.2013.817357

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. doi:10.1007/BF02294825

Millsap, R. E. (2011). *Statistical approaches to measurement invariance.* New York, NY: Routledge.

Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, *7*(1). doi:10.1186/1471-2288-7-34

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, Delta, ETS Classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics*, *32*(1), 92–109. doi:10.3102/1076998606298035

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*(3), 257–274. doi:10.1177/014662169602000306

Oliveri, M. E., Ercikan, K., & Zumbo, B. D. (2014). Effects of population heterogeneity on accuracy of DIF detection. *Applied Measurement in Education*, *27*(4), 286–300. doi:10.1080/08957347.2014.944305

Oliveri, M. E., Ercikan, K., Zumbo, B. D., & Lawless, R. (2014). Uncovering substantive patterns in student responses in international large-scale assessments—comparing a latent class to a manifest DIF approach. *International Journal of Testing*, *14*(3), 265–287. doi:10.1080/15305058.2014.891223

Paccagnella, O. (2011). Sample size and accuracy of estimates in multilevel models: New simulation results. *Methodology*, *7*(3), 111–120. doi:10.1027/1614-2241/a000029

Paek, I. (2012). A note on three statistical tests in the logistic regression DIF procedure. *Journal of Educational Measurement*, *49*(2), 121–126. doi:10.1111/j.1745-3984.2012.00164.x

Pinheiro, J. C., & Chao, E. C. (2006). Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics*, *15*(1), 58–81. doi:http://dx.doi.org/10.1198/106186006X96962

Prowker, A., & Camilli, G. (2007). Looking beyond the overall scores of NAEP assessments: Applications of generalized linear mixed modeling for exploring value-added item difficulty effects. *Journal of Educational Measurement*, *44*(1), 69–87. doi:10.1111/j.1745-3984.2007.00027.x

R Core Team. (2017). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502. doi:10.1007/BF02294403

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105–116. doi:10.1177/014662169301700201

Roussos, L. A., Schnipke, D. L., & Pashley, P. J. (1999). A generalized formula for the

Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics*, *24*(3), 293–322. doi:10.3102/10769986024003293

Schoeneberger, J. A. (2016). The impact of sample size and other factors when estimating multilevel logistic models. *The Journal of Experimental Education*, *84*(2), 373–397. doi:10.1080/00220973.2015.1027805

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159–194. doi:10.1007/BF02294572

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*(6), 613–629. doi:10.1037/0003-066X.52.6.613

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797–811. doi:10.1037//0022-3514.69.5.797

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 34, pp. 379–440). New York: Academic Press. doi:10.1016/S0065-2601(02)80009-0

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370. doi:10.1111/j.1745-3984.1990.tb00754.x

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, *27*(1), 53–75.

doi:10.3102/10769986027001053

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-Type (ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research; Evaluation, Department of National Defense. Retrieved from http://faculty.educ.ubc.ca/zumbo/DIF/

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233. doi:10.1080/15434300701375832

Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. *Journal of Educational Research & Policy Studies*, *5*(2), 1–23. Retrieved from http://eric.ed.gov/?id=EJ846827

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Olvera Astivia, O. L., & Ark, T. K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly*, *12*(1), 136–151. doi:10.1080/15434303.2014.972559

Footnotes

[1] There are a number of practical considerations, including whether to purify the matching variable with an iterative procedure that removes DIF items from the calculation of x (French & Maller, 2007), whether to use an effect size indicator to flag items that have both "statistically" and "practically" significant DIF (Hidalgo, Gomez-Benito, & Zumbo, 2014; Jodoin & Gierl, 2001; Zumbo, 1999), and the exact form of statistical test used to test $H_0 : \beta_2 = 0$ (Paek, 2012). Because there is not a single consensus view in the field regarding which choices to make for these differing options, this paper will focus on a baseline approach that uses all

items (including the studied item) when calculating x and bases the flagging of DIF items on the statistical hypothesis test that $\beta_2 = 0$.

[2] All simulation code and analysis files, including detailed tables of results for all conditions, are available upon request from the author.

[3] The default `glmer()` tolerance was used to evaluate convergence and the algorithm was restricted to attempt at most 5000 iterations.

Table 1

Type I Error Rates by Model, DIF Variance, and Sample Size.

DIF Variance	Sample Size	Model	Mean	Median	Min	Max
No	J=25, N=10	LR	0.055	0.050	0.035	0.090
		RI	0.055	0.050	0.035	0.090
		RC	0.050	0.048	0.030	0.080
	J=100, N=10	LR	0.054	0.053	0.035	0.075
		RI	0.054	0.053	0.035	0.075
		RC	0.052	0.050	0.030	0.075
	J=25, N=40	LR	0.053	0.053	0.035	0.075
		RI	0.052	0.053	0.035	0.075
		RC	0.049	0.048	0.030	0.075
	J=100, N=40	LR	0.048	0.048	0.030	0.070
		RI	0.048	0.048	0.030	0.070
		RC	0.044	0.045	0.020	0.065
Yes	J=25, N=10	LR	0.082	0.082	0.040	0.110
		RI	0.085	0.082	0.045	0.110
		RC	0.060	0.057	0.035	0.085
	J=100, N=10	LR	0.102	0.100	0.055	0.140
		RI	0.105	0.105	0.055	0.140
		RC	0.051	0.050	0.030	0.085
	J=25, N=40	LR	0.195	0.190	0.155	0.265
		RI	0.200	0.192	0.140	0.265
		RC	0.070	0.065	0.050	0.090
	J=100, N=40	LR	0.260	0.258	0.140	0.380
		RI	0.265	0.277	0.140	0.365
		RC	0.056	0.055	0.015	0.090

Note: LR=logistic regression, RI=random intercept HLR model, RC=random coefficient HLR model. Each row represents 12 conditions.

Table 2

Statistical Power by Model and DIF Variance Condition.

DIF Variance	Model	Mean	Median	Min	Max	N0	N1	N2
No	LR	0.860	0.978	0.425	1	48	36	36
	RI	0.860	0.978	0.425	1	48	36	36
	RC	0.856	0.975	0.405	1	48	36	36
Yes	LR	0.789	0.900	0.235	1	48	30	3
	RI	0.791	0.908	0.240	1	48	30	3
	RC	0.750	0.822	0.225	1	48	25	22

Note: makecell[l]LR=logistic regression, RI=random intercept HLR model, RC=random coefficient HLR model, N0=total number of conitions, N1=number of conditions with power greater than 0.8, N2=number of conditions with power greater than 0.8 and Type I error rates less than 0.08.

Table 3

DIF Results by Item.

Item		LR DIF		RI DIF		RC DIF		DIF Variance		ICC1	ICC2	Mean
11		-0.844	*	-0.853	*	-1.080	*	0.955	*	0.057	0.044	0.244
9		-0.795	*	-0.813	*	-0.813	*	0.000		0.017	0.028	0.530
20		-0.487	*	-0.491	*	-0.491	*	0.000		0.036	0.027	0.450
21		-0.294	*	-0.288		-0.288		0.000		0.031	0.027	0.428
8		-0.293	*	-0.278		-0.276		0.445	*	0.079	0.111	0.539
18		-0.247		-0.254		-0.254		0.000		0.042	0.018	0.798
2		-0.233		-0.238		-0.238		0.000		0.018	0.006	0.831
5		-0.059		-0.059		-0.058		0.014		0.000	0.000	0.565
12		-0.050		-0.034		-0.034		0.000		0.047	0.049	0.219
6		-0.033		-0.013		0.023		0.210		0.104	0.104	0.834
13		0.034		0.041		0.027		0.091		0.025	0.008	0.332
10	a	0.069		0.108		0.058		0.725	*	0.172	0.163	0.393
15		0.070		0.081						0.051	0.054	0.272
19		0.073		0.073		0.073		0.000		0.000	0.000	0.256
3		0.089		0.080						0.054	0.065	0.797
27		0.095		0.062		-0.085		0.592		0.281	0.239	0.137
22		0.143		0.139						0.080	0.069	0.461
26		0.149		0.149		0.149		0.000		0.013	0.000	0.191
17		0.158		0.156		0.156		0.000		0.059	0.040	0.295
25		0.176		0.159		0.159		0.000		0.120	0.073	0.525
4		0.202		0.224		0.229		0.036		0.060	0.095	0.639
16		0.277		0.312		0.312		0.000		0.062	0.066	0.205
14		0.347	*	0.325		0.324		0.000		0.112	0.095	0.328
7		0.350	*	0.387	*	0.350		0.450	*	0.049	0.058	0.437
23		0.432	*	0.430	*	0.338		0.488	*	0.206	0.106	0.315
24		0.506	*	0.538	*	0.538	*	0.000		0.106	0.083	0.175
1		0.539	*	0.497	*	0.708	*	1.000	*	0.195	0.159	0.789

Note: LR=logistic regression, RI=random intercept HLR model, RC=random coefficient HLR model. ICC1=unconditional ICC. ICC2=conditional ICC from RI-HLR model. Missing entries indicate items for which the DIF variance was constrained to 0 to achieve convergence. a=covariance between DIF variance and intercept variance not constrained to 0. *=p<0.05

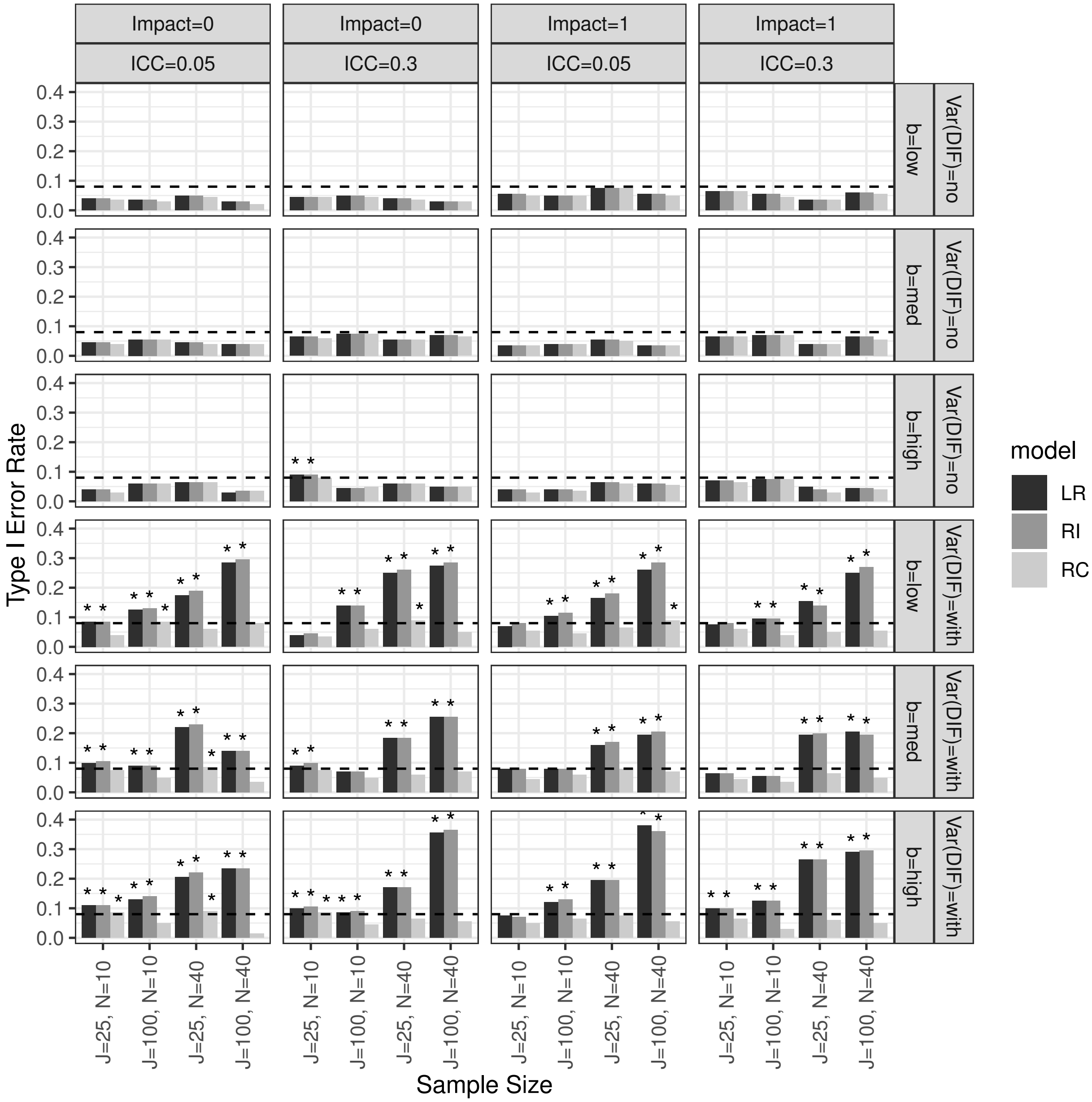

Figure 1. Type 1 Error Rates by Condition and Model. The dashed line represents y=0.08. ICC=intraclass correlation coefficient, LR=logistic regression, RI=random intercept hierarchical logistic regression, RC=random coefficient hierarchical logistic regression, low/med/high indicates difficulty of target item. *=condition in which Type I error rates are greater than 0.08.

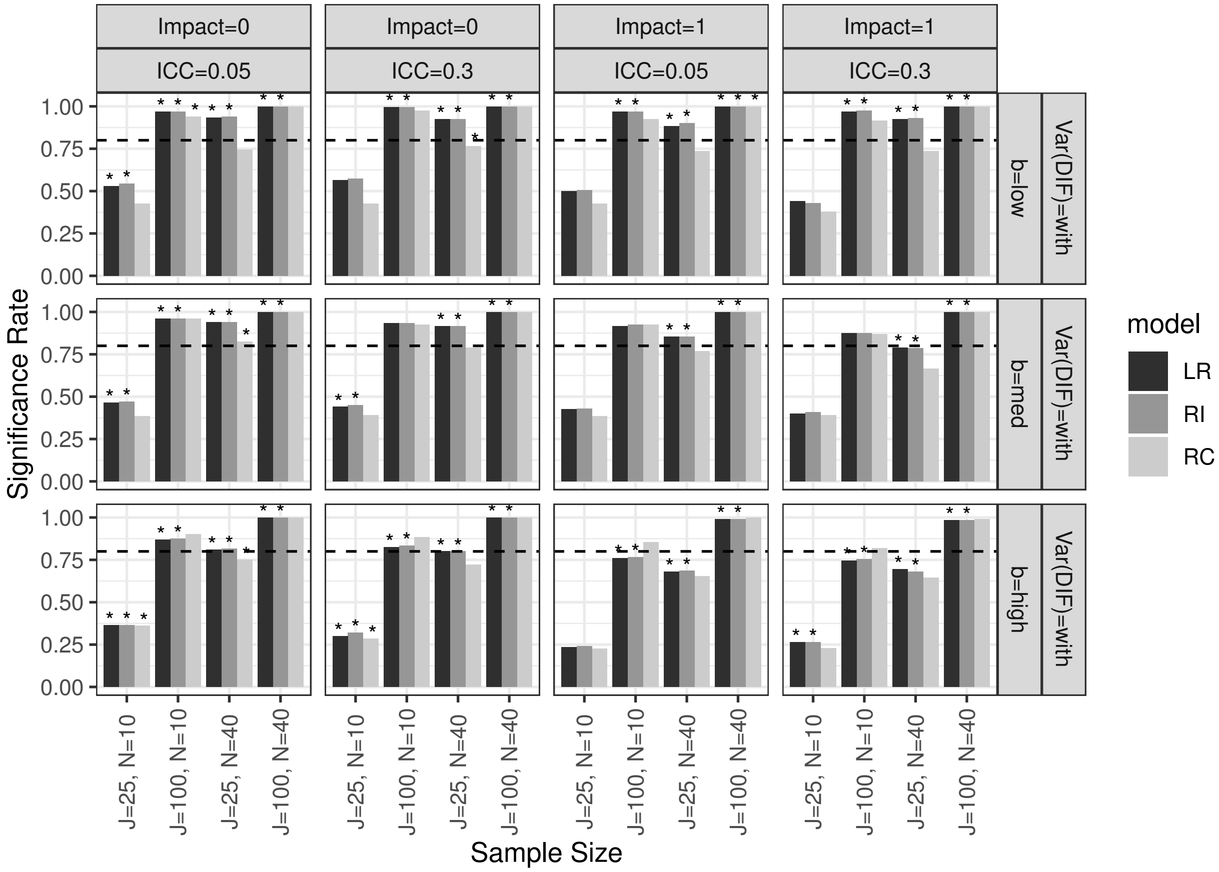

Figure 2. Statistical Power of Test for Average DIF by Condition and Model. The dashed line represents y=0.80. ICC=intraclass correlation coefficient, LR=logistic regression, RI=random intercept hierarchical logistic regression, RC=random coefficient hierarchical logistic regression, low/med/high indicates difficulty of target item. *=condition in which Type I error rates are greater than 0.08.

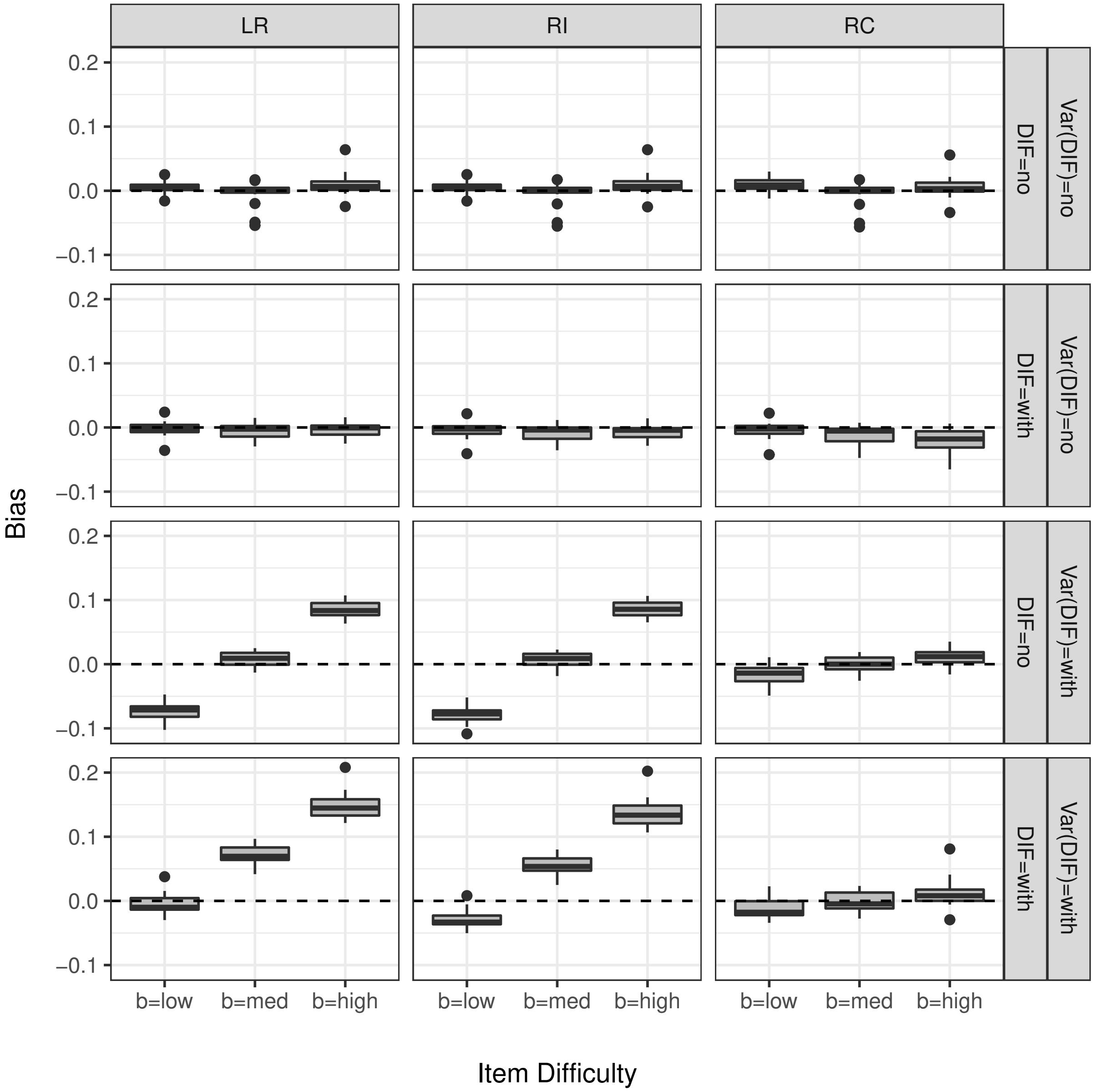

Figure 3. Bias in Estimated DIF Coefficient by DIF Condition, Item Difficulty, and Model. LR=logistic regression, RI=random intercept hierarchical logistic regression, RC=random coefficient hierarchical logistic regression, low/med/high indicates difficulty of target item. Each boxplot represents 16 simulation conditions.

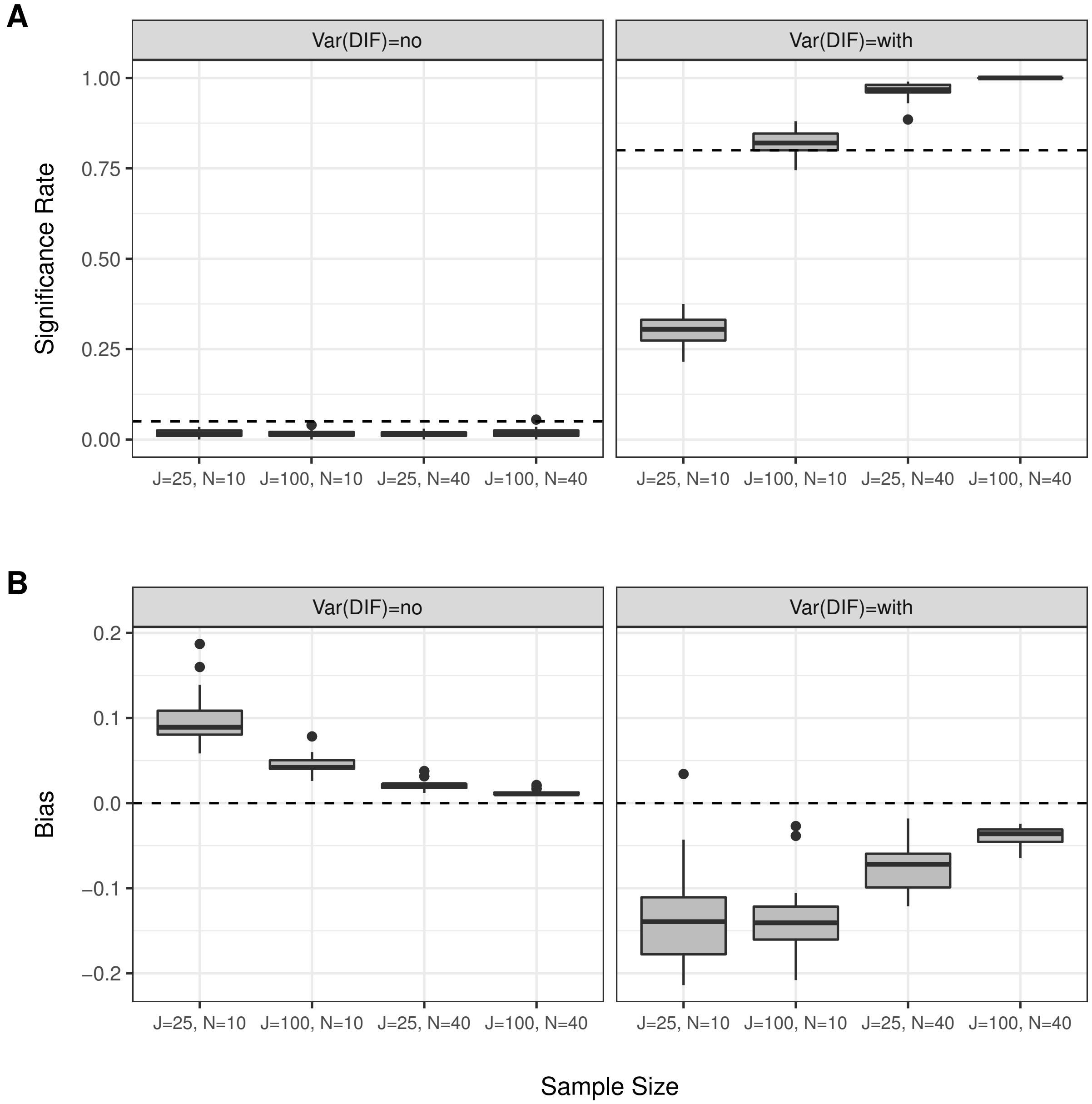

Figure 4. Significance Rates (Panel A) and Bias (Panel B) of DIF Variance Estimates by True DIF Variance and Sample Size. Each boxplot represents 48 simulation conditions. True DIF variance is 0.0 when Var(DIF)=no and 0.80 when Var(DIF)=with.