Check for updates

Original Research Article

Educational and Psychological Measurement
1–31
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00131644211068440
journals.sagepub.com/home/epm
SAGE

A Regression Discontinuity Design Framework for Controlling Selection Bias in Evaluations of Differential Item Functioning

Natalie A. Koziol[1], J. Marc Goodrich[2] and HyeonJin Yoon[1]

Abstract

Differential item functioning (DIF) is often used to examine validity evidence of alternate form test accommodations. Unfortunately, traditional approaches for evaluating DIF are prone to selection bias. This article proposes a novel DIF framework that capitalizes on regression discontinuity design analysis to control for selection bias. A simulation study was performed to compare the new framework with traditional logistic regression, with respect to Type I error and power rates of the uniform DIF test statistics and bias and root mean square error of the corresponding effect size estimators. The new framework better controlled the Type I error rate and demonstrated minimal bias but suffered from low power and lack of precision. Implications for practice are discussed.

Keywords

differential item functioning (DIF), logistic regression, regression discontinuity design, selection bias

Access to unbiased, equitable testing in education is critical to maximizing outcomes for all students (U.S. Department of Education, 2007). In modern educational models (e.g., response-to-intervention), testing is used to screen students who may be at risk

[1]University of Nebraska–Lincoln, USA
[2]Texas A&M University, College Station, USA

Corresponding Author:
Natalie A. Nebraska Center for Research on Children, Youth, Families and Schools, University of Nebraska-Lincoln, 273 Louise Pound Hall, 512 N 12th St., Lincoln, NE 68588, USA.
Email: nkoziol@unl.edu

of academic difficulties, select appropriate instructional activities, monitor student progress and responsiveness to instruction, evaluate eligibility for special education or other services (e.g., English learner services), and evaluate program effectiveness, among other purposes. Using test scores that are not adequately supported by reliability and validity evidence may have serious consequences, such as students not receiving federally mandated services for which they are eligible, or misallocation of resources away from students with the most significant educational need. Conversely, appropriate testing practices can promote inclusive educational environments and equity, and diversity in the classroom. Evaluating assessment practices to ensure they operate as intended and yield fair, unbiased outcomes is thus paramount.

Validity Evidence to Support Use of Alternate Form Test Accommodations

Assessment accommodations facilitate access to testing for diverse children with unique educational needs. According to Salvia et al. (2017), assessment accommodations can alter the way test materials are presented, the way students respond to the test, the setting in which the test takes place, and the timing of the test. One particularly common assessment accommodation is the use of alternate test forms (e.g., oral tests for children with visual impairments, translated tests for English learners). In establishing validity evidence to support the use of these alternative forms, a necessary (albeit, insufficient) step is to evaluate whether the items function in the same way (measure the same construct and are on the same scale) as the original items. Evidence to the contrary reflects differential item functioning (DIF).

Evaluating DIF is critical to supporting the use of alternate form test accommodations. For example, analyses of DIF can be performed to evaluate whether translated items function similarly to the original items (e.g., Petersen et al., 2003). However, traditional approaches for evaluating DIF are confounded by the threat of selection bias—differences between groups on variables other than the test form that was administered. An alternate form item may be more difficult, not because there is an issue with the accommodation but because the two groups differ on construct relevant (e.g., exposure to the content being tested) or irrelevant (e.g., socioeconomic status [SES]) variables. Failure to control for selection bias when evaluating DIF could result in discarding well-functioning items that are costly to develop and replace or retaining poorly functioning items that introduce bias into the testing process.

Assignment to alternate test forms is often not random. Instead, students are typically assigned based on need. For example, all students who may qualify for services as English learners must receive an English language proficiency assessment at the beginning of the school year (Lhamon & Gupta, 2015). Although there are no federally mandated standards related to assessment for English learners, best practice promotes the use of accommodations to minimize the likelihood that limited English proficiency influences performance on the assessment. One such ‘‘direct

accommodation'' is to provide the assessment in students' home language (Pitoniak et al., 2009).

In educational practice, schools and districts should rely on more information than a single screener for determining accommodations. However, Aikens et al. (2020) highlight the challenges that large-scale research studies face when determining need for accommodations as research project personnel typically do not have detailed knowledge of individual children for determining their assessment needs, and many children must be assessed within a short period of time. Consequently, the use of a single cut score, although not ideal, often represents the most feasible approach to determining accommodations in the context of large-scale research studies. Indeed, the practice of using an English language proficiency screener to determine assignment to assessment language has been used in large, federally funded survey studies, including the Early Childhood Longitudinal Study, Kindergarten Class of 1998–1999 (ECLS-K; Rock & Pollack, 2002), Kindergarten Class of 2010–2011 (ECLS-K:2011; Najarian et al., 2018), and Birth Cohort (Najarian et al., 2010). Specific information on data collection procedures for the Kindergarten Class of 2023–2024 (ECLS-K:2024) are not available, but current plans include the use of an English language screener, presumably to route multilingual children to the English or Spanish version of the assessments, as necessary (U.S. Department of Education, 2021). Other large-scale research studies that have used single indicators of English proficiency for routing children through alternate language assessments include the Head Start Family and Child Experiences Survey (FACES) and the Universal Preschool Child Outcomes Study (UPCOS; Aikens et al., 2020; Bandel et al., 2012).

The aforementioned large-scale research testing contexts naturally lend themselves to regression discontinuity design (RDD) analysis, a rigorous quasi-experimental approach for controlling selection bias when nonrandom, cut point–based assignment is used. However, with the exception of a recent application (Goodrich et al., 2021), the use of RDDs to evaluate DIF has not been considered. Given this gap in the literature, the objective of this article is twofold. First, we develop and describe two approaches for evaluating DIF within an RDD framework. Second, we use Monte Carlo simulation methods to compare the performance of these new approaches with traditional logistic regression (LR).

Methods for Investigating DIF

An item is said to exhibit DIF if the probability of a correct response for the focal group differs from that of the reference group, conditioning on the underlying latent trait (Holland & Wainer, 1993). That is, an item exhibits DIF if the group-specific item response functions (IRFs) are not perfectly overlapping. Uniform DIF reflects a group difference in difficulty or scaling, whereas nonuniform DIF reflects a group difference in discrimination (i.e., the degree to which the item differentiates among test-takers with different ability levels; Mellenberg, 1982).

Multiple approaches have been proposed for investigating DIF, including item response theory (IRT; Lord, 1980), structural equation modeling (SEM; Meredith, 1993), LR (Swaminathan & Rogers, 1990), the Mantel–Haenszel (MH) Test (Holland & Thayer, 1988), the Simultaneous Item Bias Test (SIBTEST; Shealy & Stout, 1993), and variations thereof. Broadly, these approaches differ in how they operationalize the latent trait (as a latent variable versus observed score versus corrected observed score), whether they rely on parametric assumptions, whether they allow multiple items to be tested simultaneously, and their sensitivity to nonuniform DIF. For this study, we focus on LR, as it does not require as large of a sample size as latent variable approaches, does not require coarse stratification of the matching variable, and is sensitive to both uniform and nonuniform DIF (Fidalgo et al., 2014).

Testing DIF Using LR

LR is a parametric approach for investigating DIF, specified as

$$ln\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right)=\beta_0+\beta_1\hat{\theta}_j+\beta_2 G_j+\beta_3\hat{\theta}_j G_j, \quad (1)$$

where $ln(\bullet)$ is the natural log of the odds (logit) that test-taker j correctly responds to item i, β_0 is an intercept (or threshold, in some software packages) that reflects the item's easiness (difficulty) for test-takers in group $G=0$, $\hat{\theta}_j$ is the test-taker's ability level estimated as the total test score (sum of all item responses) and β_1 reflects the item's discrimination for test-takers in group $G=0$, $\beta_0+\beta_2$ reflects the item's easiness (difficulty) for test-takers in group $G=1$, and $\beta_1+\beta_3$ reflects the item's discrimination for test-takers in group $G=1$ (Swaminathan & Rogers, 1990). Maximum likelihood is typically used for estimation. A likelihood ratio test or Wald test can be performed to test the overall hypothesis of no DIF by comparing the full model in Equation 1 to a reduced model where $\beta_2=\beta_3=0$. Under the null hypothesis, the test statistics asymptotically follow a chi-square distribution with two degrees of freedom (Paek, 2012). Alternatively, or as a follow-up to the omnibus test, one degree of freedom tests can be performed to sequentially evaluate nonuniform and uniform DIF. Nonuniform DIF is indicated if $\beta_3 \neq 0$. In the absence of nonuniform DIF, uniform DIF is indicated if $\beta_2 \neq 0$.

Two limitations of LR and related parametric observed score approaches are often cited in the DIF literature. First, $\hat{\theta}$ is subject to random and systematic error and thus groups matched on $\hat{\theta}$ may not be adequately matched on the underlying latent trait (θ). Test scores based on shorter tests and less discriminating items contain more random error, and test scores derived from items that exhibit DIF contain systematic error (DeMars, 2009, 2010; Y. Li et al., 2012; Z. Li, 2014; Rogers & Swaminathan, 1993; Shih et al., 2014). To mitigate the latter concern, a scale purification procedure (Zieky, 1993) is often recommended that involves iteratively detecting and removing all DIF items, with the exception of the item under investigation, from the calculation of the total score. Unfortunately, scale purification is a labor-intensive process

and does not always perform well (Magis & De Boeck, 2012; Shih et al., 2014). A second limitation is that Equation 1 may not adequately fit the data (DeMars, 2009, 2010; Z. Li, 2014). For example, the true IRF for multiple-choice items may have a lower asymptote due to guessing, which is captured by the IRT three-parameter logistic model (3PL) but not Equation 1. When the focal and reference groups have different underlying ability levels (i.e., when there is group ''impact,'' such that $E(\theta_j|G_j = 0) \neq E(\theta_j|G_j = 1)$), unreliability in $\hat{\theta}$ and/or incorrect specification of the functional form results in inflated Type I error rates. This inflation increases as impact and sample size increase and reliability decreases (DeMars, 2010).

Another limitation of LR is that inferences are prone to selection bias. If groups differ on variables other than the grouping mechanism and underlying latent trait, then it is unclear whether DIF is due to the grouping mechanism or some other construct relevant or irrelevant variable. Similarly, true DIF may be masked by selection bias (Wu et al., 2017).

Existing DIF Frameworks for Controlling Selection Bias

Past research has acknowledged the importance of considering selection bias in evaluations of DIF. One strategy for eliminating the threat of selection bias is to randomly assign test-takers to groups. Unfortunately, this strategy has limited utility in education, as typically the grouping mechanism either cannot be manipulated or is based on need. Two alternative strategies are to include covariates in Equation 1 (in addition to ability level; for example, Clauser et al., 1996) or apply propensity score analysis (PSA) methods (e.g., Chen et al., 2020; Liu et al., 2019). Including additional covariates is a relatively straightforward approach but assumes that all relevant covariates are measured and included in the model, and that the relationship between the covariates and item response is correctly parameterized. In the absence of random assignment, there may be numerous confounding variables that can lead to a highly parameterized model that in turn limits statistical power to detect true DIF. PSA is a diverse collection of methods that involves (a) reducing a large number of covariates into a single variable, or propensity score (i.e., balancing score), which represents the probability of being assigned to the ''treatment'' group (hereafter we use the term *treatment* to refer broadly to any grouping mechanism), given the vector of covariates, and (b) conditioning the treatment effect on the propensity score (Rosenbaum & Rubin, 1983). PSA mitigates some of the concerns with the simple covariate approach by separating the propensity score model from the treatment model and reducing the dimensionality of the covariates. Nevertheless, PSA can be complex and time-consuming and still suffers from limitations, such as the potential to overlook important covariates, a reduction in sample size and power, and sensitivity of the treatment effect to misspecification of the propensity score model. PSA does not permit inferences as strong as those of other quasi-experimental approaches, in particular RDD (Shadish & Steiner, 2010).

Testing DIF Within an RDD Framework

RDD is a quasi-experimental approach that applies when a ''running'' variable (X_j) is used to assign participants to groups (G_j) based on whether X_j exceeds a preestablished cut point (c), and interest lies in making inferences about the effect of G_j on a posttreatment outcome (Y_j) (Thistlethwaite & Campbell, 1960). Putting this in the context of a DIF investigation and drawing on a recent application (Goodrich et al., 2021), X_j could be an English proficiency screener, G_j the administration language of an achievement test where Spanish-speaking test-takers are assigned to the English form if $X_j \geq c$ and Spanish form otherwise, and Y_j the response to an item on the achievement test that is investigated for DIF.

Two alternative RDD frameworks have been developed to support causal inferences (Bloom, 2012; Cattaneo et al., 2020a, 2020b; D. S. Lee & Lemieux, 2010). The standard continuity-based framework relies on the assumption that the conditional expectations of the potential outcomes, given X_j are continuous at c, suggesting no *break* or *jump* in pretreatment factors influencing Y_j at c. This assumption ensures that no systematic differences exist between participants with similar values on X_j at c, except in terms of G_j. The nonrandom treatment assignment mechanism is completely known and statistically modeled by including X_j and G_j in the treatment model. Accordingly, G_j and Y_j are conditionally independent and the selection process is ignorable. The local randomization framework conceptualizes the RDD as a local random experiment occurring within a narrow bandwidth around the cut point. Participants near c are assumed to be identical; it is only due to random error that X_j falls slightly below or above c and thus it is only due to random error that participants are assigned to one group versus the other. Regardless of framework, the key idea is that participants in the treatment and control groups who are near the cut point are comparable on all variables other than G_j. As a result, and assuming a sharp design in which $p(G_j = 1 | X_j \geq c) = 1$ and 0 otherwise, RDD permits causal inferences on the average treatment effect *at the running variable cut point* (Bloom, 2012).

RDD treatment effects can be estimated using graphical, parametric, or nonparametric methods. In our proposed framework, we focus on nonparametric methods, based on the recommendations of Cattaneo et al. (2020b) who advise against parametric methods. We first propose using local linear regression within an RDD continuity-based framework (hereafter abbreviated as LLn-RDD) to test for DIF. This approach entails fitting the following weighted least-squares regression:

$$Y_{ij} = \beta_0 + \beta_1 g_j + \beta_2 (X_j - c) + \beta_3 g_j (X_j - c), \tag{2}$$

where the target parameter is given by β_1 and reflects the magnitude of uniform DIF at c. Calculation of weights (w_{ij}) depends on the chosen kernel function and bandwidth (h_i). We recommend a triangular kernel function (Cattaneo et al., 2020b):

$$w_{ij} = \begin{cases} 1 - \left| \frac{X_j - c}{h_i} \right| \ if \left| X_j - c \right| \leq h_i \\ 0 \ if \left| X_j - c \right| > h_i \end{cases}. \tag{3}$$

Equation 3 highlights the fact that only participants with X_j sufficiently close to c (as defined by h_i) contribute to the estimation of the treatment effect. It is these cases that define the ''effective'' sample size. The optimal bandwidth is one that supports the linear approximation between $X_j - c$ and Y_{ij} imposed by Equation 2 (i.e., minimizes bias of the treatment effect estimator) while it minimizes the variance of the treatment effect estimator. We recommend a bandwidth that minimizes the mean square error (MSE) of the treatment effect estimator (i.e., the MSE-optimal bandwidth; Cattaneo et al., 2020b). However, selecting the MSE-optimal bandwidth concedes that misspecification error is not zero. Consequently, standard ordinary least-squares (OLS) standard errors and confidence intervals, which assume no misspecification error, are inappropriate. Robust bias-corrected standard errors and confidence intervals are instead recommended. Data-driven approaches for selecting bandwidths and robust bias-corrected inference are automatized by RDD software packages and interested readers can refer to Cattaneo et al. (2020b) for formulas and their theoretical foundation.

Local linear regression is often used on categorical outcomes as it does not require that Y_{ij} follow a normal distribution or that the global association between X_j and Y_{ij} is linear, only that the local association between X_j and Y_{ij} is approximately linear. However, nonparametric local logit RDD estimation (hereafter abbreviated as LLg-RDD; Xu, 2017) in which the local polynomial approximation is performed on the logit scale rather than the probability scale, may be preferable. Derivations for an asymptotic MSE (AMSE) optimal bandwidth and corresponding robust bias-corrected standard errors and confidence intervals, and justification for a uniform kernel function, are given by Xu (2017).

A linear approximation is likely to be supported across a broader range of the outcome when applied to the logit scale, thereby permitting broader bandwidths and larger sample sizes. This could result in greater precision. For example, outside of the RDD and DIF contexts, Frölich (2006) found that local logit estimators had greater precision than local linear estimators for dichotomous outcomes with many regressors. On the contrary, outside of the DIF context, Xu (2017) observed limitations with the *ASME* optimal bandwidth and noted that standard errors were large, suggesting that power may suffer.

There are several noteworthy differences between the traditional and proposed approaches for testing DIF. First, LR permits tests of uniform and nonuniform DIF, whereas LLn-RDD and LLg-RDD as defined above are limited to tests of uniform DIF. Testing whether the item's discrimination varies between groups would require either imposing parametric assumptions (thereby increasing susceptibility to bias) or subsetting the analyses along discrete levels of θ (cf. Mazor et al., 1994; thereby decreasing power). Second, LR attempts to control for θ_j by including $\hat{\theta}_j$ as a covariate, whereas the inclusion of $\hat{\theta}_j$ is not necessary in LLn-RDD or LLg-RDD. This follows from the continuity assumption that ensures no jump in the association between *pretreatment* covariates and Y_j at c. In the testing contexts applicable to our proposed approach (i.e., where G_j represents alternative test forms), θ_j is a pretreatment

covariate; assigning test-takers to different forms does not change their underlying ability level, only potentially their observed score. It is possible to include $\hat{\theta}_j$ as a covariate in Equation 2 as a means for increasing precision. The concern is that $\hat{\theta}_j$ may not provide a good approximation of θ_j and may be impacted by G_j, such that the covariate-adjusted RDD estimator would not be a consistent estimator of the average effect at c (Cattaneo et al., 2020b). Third, LLn-RDD and LLg-RDD estimates of DIF generalize to test-takers in the population with $X_j = c$, whereas LR inferences are not conditional on X_j. Finally, LR and LLg-RDD attend to the bounded and categorical nature of Y_{ij} by modeling the logit of a correct response, whereas LLn-RDD predicts Y_{ij} directly.

Taken as a whole, the RDD approaches have both advantages and disadvantages when compared with LR for detecting DIF. Their advantages are that they control for selection bias, use nonparametric methods which require fewer assumptions and are more robust to outliers and idiosyncrasies in the data that are far from c, and do not require estimation of θ_j. Their disadvantages are that they are limited to testing uniform DIF, inferences are limited to a small fraction of the total population (i.e., test-takers with $X_j = c$), and, in most cases, they are likely to have lower power due to the effective sample size being smaller than the total sample size.

The Current Study

Although the RDD approaches have some theoretical advantages for evaluating uniform DIF, it is unclear how these approaches perform in practice when sample conditions are less than ideal. Empirical evidence is needed to support their use. The purpose of this Monte Carlo simulation study was to compare the performance of LR, LLn-RDD, and LLg-RDD in detecting the absence, presence, and magnitude of uniform DIF across varying sample conditions, including different magnitudes of group impact, magnitudes of selection bias, sample sizes, test lengths, and item properties. Four research questions were posed as follows:

Research Question 1 (RQ1): How does the Type I error rate of the LR, LLn-RDD, and LLg-RDD uniform DIF test statistics compare across varying sample conditions?

Research Question 2 (RQ2): How does the power of the LR, LLn-RDD, and LLg-RDD uniform DIF test statistics compare across varying sample conditions?

Research Question 3 (RQ3): How does the bias of the LR, LLn-RDD, and LLg-RDD uniform DIF effect size estimators compare across varying sample conditions?

Research Question 4 (RQ4): How does the root mean square error (RMSE) of the LR, LLn-RDD, and LLg-RDD uniform DIF effect size estimators compare across varying sample conditions?

Based on prior research, we hypothesized that the LR DIF test statistic would demonstrate inflated Type I error rates and the effect size estimator would be biased when the magnitude of impact was large and the test was short, and in the presence of selection bias, particularly when sample size was large and the target item was strongly discriminating (DeMars, 2009, 2010; Y. Li et al., 2012; Liu et al., 2019; Rogers & Swaminathan, 1993; Shih et al., 2014). We expected the corresponding LLn-RDD and LLg-RDD test statistics and effect size estimators would be robust to selection bias, group impact, and test length. Controlling for differences in the Type I error rate and bias, we hypothesized that the LR approach would be more powerful and precise than the RDD approaches, and that the LLg-RDD approach would be more precise than the LLn-RDD approach (Frölich, 2006).

This study focuses on uniform DIF because it is a natural starting point for evaluating the RDD approaches. These approaches are not designed to detect interactions with continuous variables (in this case, the proficiency by group interaction reflecting nonuniform DIF). If they do not perform well for detecting uniform DIF then they are even less likely to perform well for detecting nonuniform DIF. We acknowledge in the ''Discussion'' section, however, that investigating nonuniform DIF is an important future direction.

Method

Design

Five simulation factors were fully crossed for a total of 216 conditions: (a) Group impact (three levels), (b) Selection bias (three levels), (c) Sample size (three levels), (d) Test length (two levels), and (e) Item properties (four levels). $R = 1{,}050$ replications were generated for each combination of impact, selection bias, sample size, and test length for a total of 56,700 replications. Item properties were varied within replications (i.e., each simulated test contained all combinations of items). Within each condition, only the first 1,000 replications for which all three analyses' approaches converged were used to evaluate the test statistics and effect size estimators.

Group Impact. The levels of group difference in true proficiency were 0 *SD*, .5 *SD*, and 1 *SD*, representing no mean impact, moderate impact, and large impact, respectively. This range mirrors levels considered in prior research (e.g., DeMars, 2009; Hidalgo et al., 2014; Y. Li et al., 2012; Narayanan & Swaminathan, 1996).

Selection Bias. For the target items, the probability of a correct response was generated to be a function of the traditional 3PL IRT item and person properties, in addition to a person-level confounding variable, the RDD running variable. This variable was generated to account for no, minimal, or moderate variability in the item responses (see ''Data Generation'' section).

Sample Size. Three sample sizes were considered: $n_r = n_f = 150$ ($N = 300$), $n_r = n_f = 300$ ($N = 600$), and $n_r = n_f = 1{,}000$ ($N = 2{,}000$). Whereas unequal sample sizes are more likely to be observed in practice, we imposed the simplifying assumption of equal sample sizes to prevent confounding variability (variability between sample size conditions due to factors other than sample size) that could potentially arise from generating unequal sample sizes.[1] We acknowledge this limitation in the ''Discussion'' section. The smallest sample size condition falls below ETS' minimum recommended total sample size of 500 and group sample size of 200 during the test assembly phase (Zwick, 2012), but represents a plausible sample size when considering special populations such as English language learners or students with disabilities. For example, only 150 students enrolled in the ECLS-K:2011 completed the Spanish spring kindergarten mathematics assessment (Najarian et al., 2018). The middle sample size condition meets minimum guidelines but is still relatively small, whereas the largest sample size represents an ideal scenario and is similar to the largest condition considered in prior research (e.g., Jodoin & Gierl, 2001; Y. Li et al., 2012). Practitioners may not have access to a sample size of 1,000, particularly for the focal group, when DIF analyses are not planned/powered a priori. We include this largest condition to help inform sample size planning for DIF analyses when sample size is under the control of the practitioner.

Test Length. Short (20 items) and long (80 items) tests were generated. Twenty items has been recommended as a lower bound for investigating DIF (Zumbo, 1999). Although short, 20-item tests are used in practice (e.g., the ECLS-K:2011 kindergarten science achievement test; Najarian et al., 2018). Past simulation research has considered 80 items to represent a long test, and similar test lengths are used in practice (e.g., the ECLS-K:2011 kindergarten reading and mathematics achievement tests; Najarian et al., 2018).

Item Properties. Four combinations of item discriminations and difficulties were considered for the target items: (a) high discrimination ($a = 1.6$), low difficulty ($b = -1.5$); (b) low discrimination ($a = 0.6$), moderate difficulty ($b = 0.0$); (c) high discrimination ($a = 1.6$), moderate difficulty ($b = 0.0$); and (d) high discrimination ($a = 1.6$), high difficulty ($b = 1.5$). These combinations of items have been investigated in prior DIF research (Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993) and were chosen for this study because they contribute varying information and target different locations across the latent trait continuum.

Data Generation

To help with interpretation, we use the applied example of Goodrich et al. (2021) to describe the simulated testing context. That is, we consider a scenario in which Spanish-speaking kindergarteners are administered a mathematics assessment and the language of administration is determined based on their performance on an English

language screener. Following a sharp RDD design, all students who pass the English proficiency cutoff are administered the mathematics assessment in English (reference group) and all students who do not pass are administered the assessment in Spanish (focal group).

Item responses were generated in base R Version 3.6.1 (R Core Team, 2019) according to a modified 3PL IRT model:

$$p_{ij,g}\left(Y_{ij,g}=1|X_j,\theta_{j,g},a_i,b_{i,g},c_i,\gamma\right)=c_i+\frac{1-c_i}{1+e^{-a_i\left(\theta_{j,g}-b_{i,g}-\gamma_i X_j^3\right)}}. \quad (4)$$

Notation is as follows: $p_{ij,g}$ is the probability of a correct response to mathematics item i ($i=1,\ldots,L$; $L\in[20,80]$) for kindergartener j ($j=1,\ldots,n_g$; $n_g\in[150, 300, 1,000]$) assigned to mathematics assessment form G where $G=0$ (Spanish form) or 1 (English form); X_j is an English language screener (the RDD running variable) used to determine the mathematics assessment language: $G=0$ if $X_j<0$, otherwise $G=1$; θ_j is the kindergartener's latent mathematics ability (the distributions of X_j and θ_j are detailed below); a_i, $b_{i,g}$, and c_i are item discrimination, difficulty, and pseudo-guessing parameters (detailed below); and γ_i is the confounding effect of English language proficiency (detailed below). Item responses were generated by comparing the probability of a correct response with a random number generated from a uniform(0, 1) distribution.

Impact was simulated by generating X_j and θ_j to follow a bivariate normal distribution: $\begin{bmatrix}X_j\\ \theta_j\end{bmatrix}\sim N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & r\\ r & 1\end{bmatrix}\right)$, where $r\in[0, .313, .628]$, such that $\mu_{\theta_g}=(-1)\times d/2+G\times d$ and $d\in[0, .5, 1]$. This approach is consistent with the continuity assumption underlying RDD as $\mu_{\theta_{g=0}}=\mu_{\theta_{g=1}}$ at c regardless of r.

For both test length conditions, eight items (four non-DIF and four uniform DIF) were targeted for investigation (see Table 1). The properties of the four non-DIF items match those described in the ''Item Properties'' section. The discrimination parameters of the four DIF items were the same as those of the non-DIF items, whereas the difficulty parameters were chosen, such that the area between the IRFs of the two groups was equal to .6 (reflecting a moderate level of DIF; Swaminathan & Rogers, 1990) and the group-specific difficulty parameters were equidistant from the target difficulty parameter. Given these constraints, the item difficulties were derived by solving the following equation that quantifies the area between two response functions under the assumption that $a_{i,g=0}=a_{i,g=1}$ (S. Lee, 2017):

$$Area=(1-c_i)\left|b_{i,g=0}-b_{i,g=1}\right|. \quad (5)$$

DIF was generated to be unidirectional, so DIF items were always easier for the reference group.

The target items accounted for 40% of the 20-item test. To ensure similar item properties and maintain a constant proportion (.20; see Gierl et al., 2004) of DIF

Table 1. Generating IRT Properties and Observed Classical Test Theory Properties of Target Items

Item	a_i	$b_{i,g=0}$	$b_{i,g=1}$	c_i	$\bar{p}_i$	$\bar{\rho}_{y_i,\theta}$
1	1.6	−1.5	−1.5	.2	.87	.39
2	0.6	0	0	.2	.60	.23
3	1.6	0	0	.2	.60	.45
4	1.6	1.5	1.5	.2	.33	.28
5	1.6	−1.125	−1.875	.2	.86	.41
6	0.6	0.375	−0.375	.2	.60	.25
7	1.6	0.375	−0.375	.2	.60	.45
8	1.6	1.875	1.125	.2	.34	.30

Note. IRT = item response theory; a_i = item discrimination; $b_{i,g=0}$ = item difficulty for focal group; $b_{i,g=1}$ = item difficulty for reference group; c_i = item pseudo-guessing parameter; $\bar{p}_i$ = observed proportion of correct responses averaged across replications and conditions; $\bar{\rho}_{y_i,\theta}$ = observed point-biserial correlation between the item and latent trait score averaged across replications and conditions.

items across test lengths, the eight target items were replicated 4 times for the 80-item test. Properties of the remaining 60% of items (i.e., the remaining 12 items of the 20-item test and 48 items of the 80-item test) were randomly generated for each replication under the following constraints: $a_i \sim lognormal(0, .1225)$ and $b_i \sim N(0, 1)$ with b_i truncated at [–2, 2] and $b_{i,g=0} = b_{i,g=1}$ (DeMars, 2009; Magis & De Boeck, 2012). For all items, $c_i = .2$.

To simulate selection bias, it was necessary to generate a variable besides the mathematics ability variable that was related to both group membership (mathematics assessment language) and item response. The English language screener, by definition under the RDD, predicted group membership. As noted above, the 3PL IRT model was modified so that the English language screener also predicted response to the target items. The relationship between the screener and outcome was chosen to be nonlinear to ensure that a narrower bandwidth would be necessary under the RDD approaches. Three magnitudes of effects were considered: $\gamma_i = 0$, -0.04, and -0.10, corresponding to no selection bias, minimal bias, or moderate bias, respectively. For the nontarget items, γ_i was fixed at 0.

Note that generating $\gamma_i \neq 0$ is akin to generating another source of DIF, DIF that is due to English language proficiency (a student characteristic) rather than G (the test form that was administered to the student). In our hypothetical context, for example, it might be the case that word problems (more language-intensive items) exhibit DIF due to language proficiency.

Data Analysis

LR, LLn-RDD, and LLg-RDD were used to investigate DIF. The LR approach, specified according to Equation 1 but without the ability by group term, was carried out

in Mplus Version 8.5 (Muthén & Muthén, 1998–2020), using maximum likelihood estimation. An item was flagged as DIF if the Wald test for the group effect was significantly different from 0 ($p < .05$). The LLn-RDD approach was implemented within the rdrobust package in *R* (Calonico et al., 2021) according to Equation 2. Bandwidths were empirically derived based on a triangular kernel function and MSE-optimal bandwidth selector. Estimation was carried out using OLS but with robust bias-corrected standard errors. An item was flagged for DIF if $p < .05$ for the group difference. LLg-RDD was implemented within the rd.categorical package in R (Xu, 2017). Bandwidths were derived from the AMSE-optimal bandwidth selector with a uniform kernel function.

Monahan et al. (2007) describe several effect sizes appropriate for quantifying the magnitude of uniform DIF. For this study, effect size was measured as the group difference in the predicted proportion of respondents with a correct response (p_{DIF}) as this effect size can be approximated by all three DIF approaches. For both RDD approaches, the estimated group difference is on the proportion scale, so no additional calculations were required. For the LR approach, the effect size was calculated using the conditional-difference-in-proportions definition (Monahan et al., 2007):

$$\text{LR-STD-P-DIF} = \frac{\sum_m w_m \left(P_{rm}^{LR} - P_{fm}^{LR} \right)}{\sum_m w_m} \tag{6}$$

where m is defined by the range of scores observed on the matching criterion (mathematics sum score), w_m is a weight equal to the total number of kindergarteners with a mathematics sum score equal to m, and P_{rm}^{LR} and P_{fm}^{LR} are the model-predicted proportions of a correct response for kindergarteners in the reference and focal groups, respectively, who achieved a mathematics sum score equal to m.

For LR, a purification procedure was performed in which the mathematics score used as the matching criterion was calculated as the sum of only the responses to the non-DIF items plus the item under investigation (Zieky, 1993). Because this procedure was not under investigation, purification was based on truth (DIF items were treated as known) as opposed to carried out using an estimative iterative procedure. This approach thus presents a best-case scenario.

Outcomes

The proportion of converged replications (out of 1,050) was documented for the three approaches. For each of the four target non-DIF items, the Monte Carlo estimated Type I error rate was calculated as the proportion of the first 1,000 converged replications that the item was incorrectly flagged as DIF. Using a normal approximation to the binomial, it is expected with 99% confidence that a test statistic with a true Type I error rate of .05 will have an estimated error rate between .032 and .068. Similarly, for each of the four target DIF items, power was calculated as the proportion of the first 1,000 converged replications that the item was correctly flagged as DIF. Power

was only interpreted when the corresponding Type I error rate did not fall outside the 99% confidence bounds.

For all target items, bias of the effect size estimator was calculated as the average of effect size estimates across the first 1,000 converged replications minus the true effect size: $\sum_r \hat{p}_{DIF,r}/R - p_{DIF}$. For non-DIF items, the true effect size was 0. For DIF items, the true effect size was approximated at the English language proficiency cutoff, using an IRT model-based standardization similar to Equation 6 but involving numerical integration over the true mathematics score (θ) instead of summation over the observed scores, and using the generating IRT parameters to obtain $P_{r\theta}$ and $P_{f\theta}$. The p index was .16 (Category C; Monahan et al., 2007) for the item with high discrimination and moderate difficulty and .08 to .09 (Category B; Monahan et al., 2007) for the other items. For each estimate of bias, a 99% confidence interval was calculated to determine whether bias was significantly different from 0. For DIF items in which the true effect size was not 0, relative bias was calculated by dividing the estimated bias by the true effect size. RMSE of the effect size estimator was calculated as $\sqrt{\widehat{bias}(\hat{p}_{DIF})^2 + \widehat{Var}(\hat{p}_{DIF})}$ where $\widehat{Var}(\hat{p}_{DIF}) = \sum_r (\hat{p}_{DIF,r} - \sum_r \hat{p}_{DIF,r}/R)^2/R$. Similar to power, RMSE was only interpreted when the corresponding estimator was not significantly biased.

Given the large number of conditions, an analysis of variance (ANOVA) was performed on the aggregated Type I error and bias data to identify which simulation factors accounted for a meaningful proportion of variability in the outcomes. Interpretation was limited to effects with $\eta^2 \geq .02$ (Cohen's, 1988, cutoff for a small effect). Visual inspection was performed for the power and RMSE outcomes in lieu of ANOVA due to data missing not at random (power and RMSE data were omitted if the corresponding Type I error rate and bias were unacceptable).

Results

The primary results are organized below by outcome. We first summarize key characteristics of the data generation and analysis conditions to contextualize the primary results.

Classical test theory properties of the target items, averaged across replications and conditions, are shown in Table 1. As expected, the proportion of correct responses was highest for the low difficulty items ($\bar{p}_i = .86 - .87$), in the middle for the moderate difficulty items ($\bar{p}_i = .60$), and lowest for the high difficulty items ($\bar{p}_i = .33 - .34$). The point-biserial correlation between the items and latent trait scores was higher for the high discrimination, low difficulty items ($\bar{\rho}_{y_i,\theta} = .39 - .41$) and high discrimination, moderate difficulty items ($\bar{\rho}_{y_i,\theta} = .45$) than the low discrimination, moderate difficulty items ($\bar{\rho}_{y_i,\theta} = .23 - .25$) and high discrimination, high difficulty items ($\bar{\rho}_{y_i,\theta} = .28 - .30$). Differences in $\bar{\rho}_{y_i,\theta}$ across the high discrimination items are due to differences in the distance between the items' difficulty and the sample's ability level, as well as the inclusion of a lower asymptote in the generating IRT model that impacts the location at which the items provide maximal information.

The percentage of converged replications was 100% across all conditions for LR and LLn-RDD. For LLg-RDD, convergence was less than 100% (ranging from 97.3% to 99.9% with a median of 99.6%) for 19 of the 54 conditions. Among these conditions, greater rates of nonconvergence were observed for the large impact and small sample size conditions. The effective sample size ranged from 47% to 54% of the total sample size for LLn-RDD and 57% to 83% for LLg-RDD. LR analyses were based on data from the full sample.

Type I Error

The Monte Carlo estimated Type I error rates are illustrated in Figure 1 and complete numerical results are available in Table S1. In the figure, Type I error rate is indicated by the *x*-axis with dashed vertical lines, indicating 99% confidence bounds for a true Type I error rate of .05; sample size and impact conditions are represented by columns, test length, and selection bias; item property conditions are represented by rows; and DIF approach is indicated by different symbols (plus = LR, circle = LLn-RDD, triangle = LLg-RDD).

The observed Type I error rates were more variable across conditions, and more inflated on average, under the LR approach (M = .15, range = .04–.93) than the LLn-RDD (M = .06, range = .04–.09) and LLg-RDD (M = .04, range = .02–.06) approaches. Results from an ANOVA identified a four-way interaction, selection bias by DIF method by sample size by item (η^2 = .02), that accounted for a meaningful proportion of variability in Type I error rates. Impact and test length did not account for a meaningful proportion of variability. The LR Type I error rate was more inflated when selection bias was present, and this pattern was more pronounced when sample size was large and for the two items with high item-ability correlations (the high discrimination, low difficulty and high discrimination, moderate difficulty items). The LLn-RDD and LLg-RDD Type I error rates were not sensitive to selection bias, sample size, or item properties.

Power

The Monte Carlo estimated power rates are shown in Figure 2 for the conditions in which the corresponding estimated Type I error rate did not exceed the 99% confidence bounds for a true Type I error rate of .05. Numerical results are provided in Table S2. The figure follows the same structure as before but with power on the *x*-axis. Power to detect DIF was consistently higher for the LR approach than the LLn-RDD and LLg approaches, with an average difference in power of .56 (range = .06–.80) and .54 (range = .09–.85), respectively. Power was slightly higher on average for LLn-RDD than LLg-RDD (M_{Diff} = .05, range = −.03–.20). Even under the largest sample size condition, power of the LLn-RDD and LLg-RDD test statistics did not reach .80. In contrast, power of the LR test statistic exceeded .80 under the smallest sample size condition for the two items with high item-ability correlation. For all

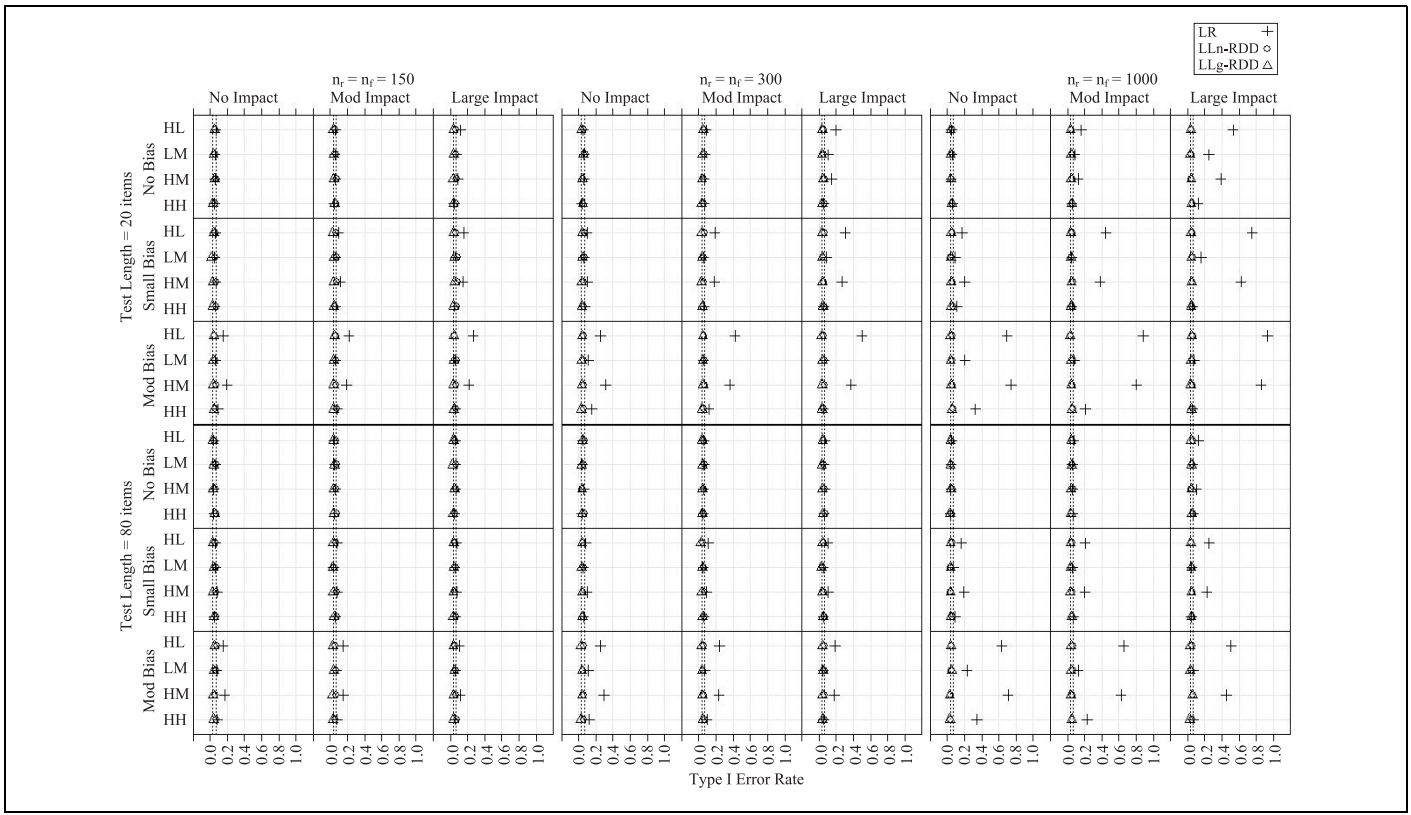

Figure 1. Monte Carlo Estimated Type I Error Rate

Note. Dashed vertical lines indicate 99% confidence bounds (.032, .068) for a true Type I error rate of .05. LR = logistic regression; LLn-RDD = local linear regression, regression discontinuity design; LLg-RDD = local logit estimation, regression discontinuity design; n_r = sample size of reference group; n_f = sample size of focal group; Impact = latent mean group difference (0 *SD*, .5 *SD* [Mod], and 1 *SD* [Large]); Bias = selection bias; HL = item with high discrimination (a = 1.6), low difficulty (b = −1.5); LM = item with low discrimination (a = 0.6), moderate difficulty (b = 0.0); HM = item with high discrimination (a = 1.6), moderate difficulty (b = 0.0); HH = item with high discrimination (a = 1.6), high difficulty (b = 1.5).

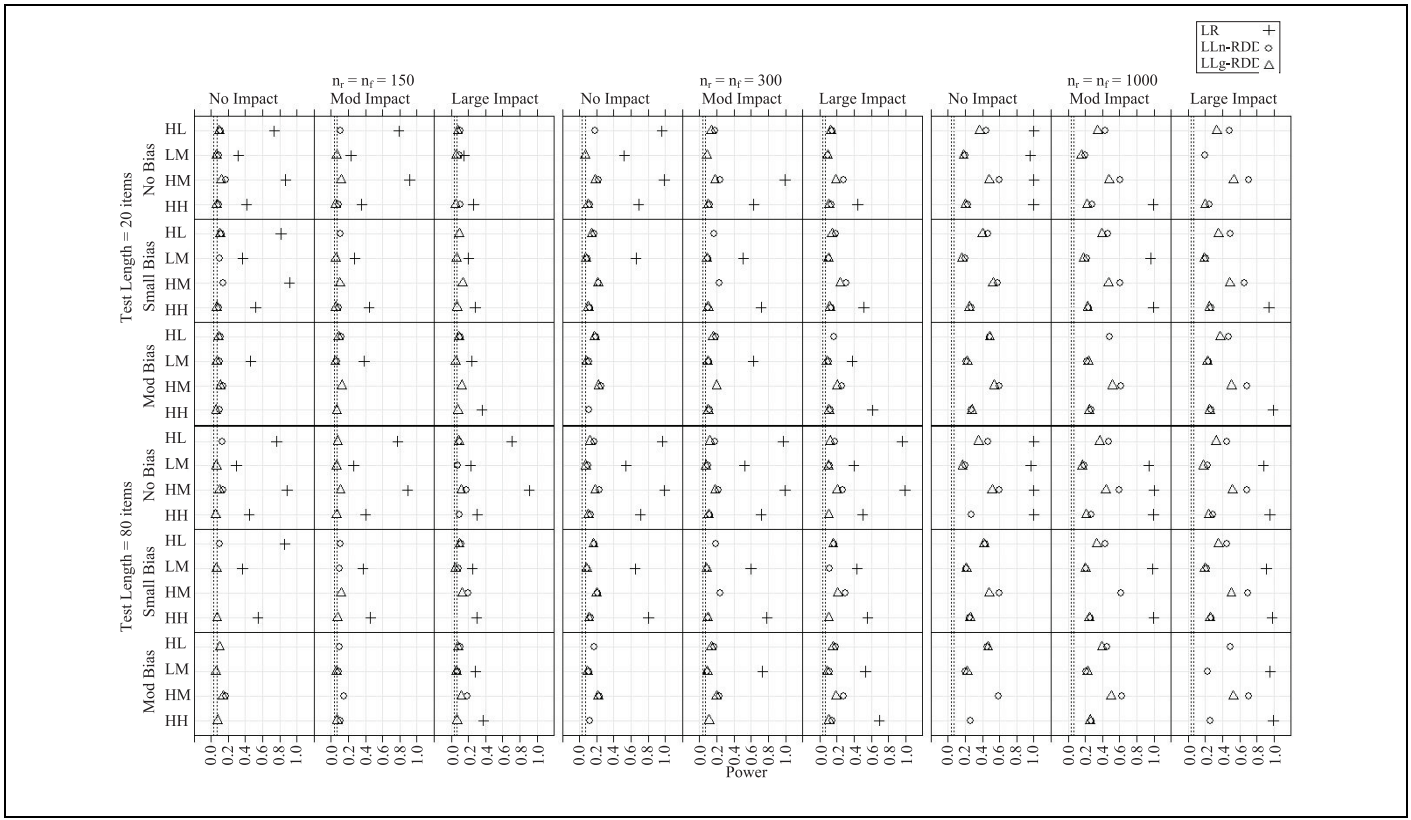

Figure 2. Monte Carlo Estimated Power

Note. Power is only displayed where the corresponding estimated Type I error rate does not exceed the 99% confidence bounds for a true Type I error rate of .05. Dashed vertical lines indicate 99% confidence bounds (.032, .068) for a true Type I error rate of .05. LR = logistic regression; LLn-RDD = local linear regression, regression discontinuity design; LLg-RDD = local logit estimation, regression discontinuity design; n_r = sample size of reference group; n_f = sample size of focal group; Impact = latent mean group difference (0 *SD*, .5 *SD* [Mod], 1 *SD* [Large]); Bias = selection bias; HL = item with high discrimination (a = 1.6), low difficulty (b = −1.5); LM = item with low discrimination (a = 0.6), moderate difficulty (b = 0.0); HM = item with high discrimination (a = 1.6), moderate difficulty (b = 0.0); HH = item with high discrimination (a = 1.6), high difficulty (b = 1.5).

three approaches, power was higher for the two items with high item-ability correlations and when sample size was large.

Bias

Monte Carlo estimated bias of p_{DIF} is illustrated in Figure 3 (Table S3) for the non-DIF items and Figure 4 (Table S4) for the DIF items. The *x*-axis indicates bias on the p_{DIF} (probability) scale, where the dashed vertical line indicates an optimal value of 0. For the non-DIF items, bias ranged from −0.03 to 0.06 (M = 0.02) under LR, −0.01 to 0.02 (M = 0.00) under LLn-RDD, and −0.02 to 0.01 (M = −0.01) under LLg-RDD. For the DIF items, bias ranged from −0.03 to 0.05 (M = 0.01) under LR, −0.02 to 0.03 (M = 0.00) under LLn-RDD, and −0.03 to 0.02 (M = −0.01) under LLg-RDD.

Results from the two ANOVAs revealed similar patterns across the non-DIF and DIF items. The method by impact by item interaction accounted for a meaningful proportion of variability in bias (η^2 = .05 and .06 for the non-DIF and DIF items, respectively). When impact was small, the LR p_{DIF} estimator demonstrated similar levels of bias across items. When impact was large, the pattern diverged; bias became more positive for the two items with high item-ability correlations and more negative for the two items with low item-ability correlations. Under the LLn-RDD and LLg-RDD approaches, bias was less variable and closer to zero across items and impact levels, particularly for the non-DIF items. There was also a meaningful method by selection bias interaction (η^2 = .19 and .17 for the non-DIF and DIF items, respectively). The LR p_{DIF} estimator, and to a lesser extent the LLg-RDD estimator (apparent under the small sample size condition) became more biased as selection bias increased, whereas the LLn-RDD estimator was not sensitive to selection bias.

Root Mean Square Error

Monte Carlo estimated RMSE of p_{DIF} is shown in Figures 5 and 6 (Tables S5 and S6) for the non-DIF and DIF items, respectively. RMSE is only displayed when the p_{DIF} estimator was not significantly biased. The *x*-axis indicates RMSE on the p_{DIF} (probability) scale, where the dashed vertical line indicates an optimal value of 0. Among the conditions in which the p_{DIF} estimator was not significantly biased, the LR estimator was consistently more precise than the LLn-RDD and LLg-RDD estimators, with an average difference in RMSE of .09 (range = .03–.14) and .05 (range = .02–.09), respectively, for the non-DIF items and an average difference in RMSE of .11 (range = .05–.15) and .06 (range = .03–.09), respectively, for the DIF items. LLg-RDD was more precise than LLn-RDD, with an average difference in RMSE of .04 (range = .01–.07) for the non-DIF and DIF items. RMSE reached as high as .20 under the LLn-RRD approach, with an average value of .12 and minimum of .04. For all approaches, greatest precision was observed for the item with high discrimination and low difficulty and when sample size was large.

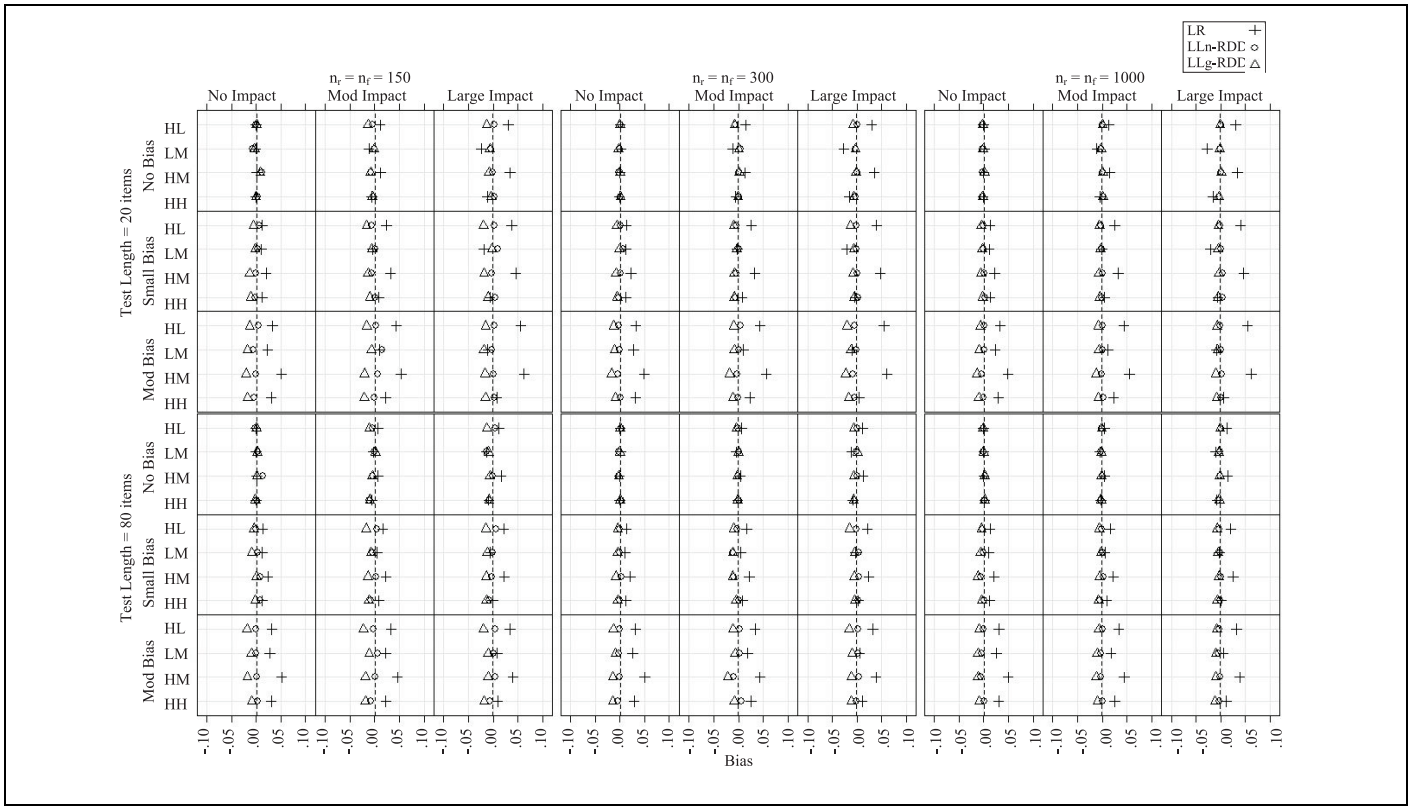

Figure 3. Monte Carlo Estimated Bias of p_{DIF} for Non-DIF Items

Note. Dashed vertical line indicates where bias = 0. DIF = differential item functioning; LR = logistic regression; LLn-RDD = local linear regression, regression discontinuity design; LLg-RDD = local logit estimation, regression discontinuity design; n_r = sample size of reference group; n_f = sample size of focal group; n_r = sample size of reference group; n_f = sample size of focal group; Impact = latent mean group difference (0 *SD*, .5 *SD* [Mod], 1 *SD* [Large]); Bias = selection bias; HL = item with high discrimination (a = 1.6), low difficulty (b = −1.5); LM = item with low discrimination (a = 0.6), moderate difficulty (b = 0.0); HM = item with high discrimination (a = 1.6), moderate difficulty (b = 0.0); HH = item with high discrimination (a = 1.6), high difficulty (b = 1.5).

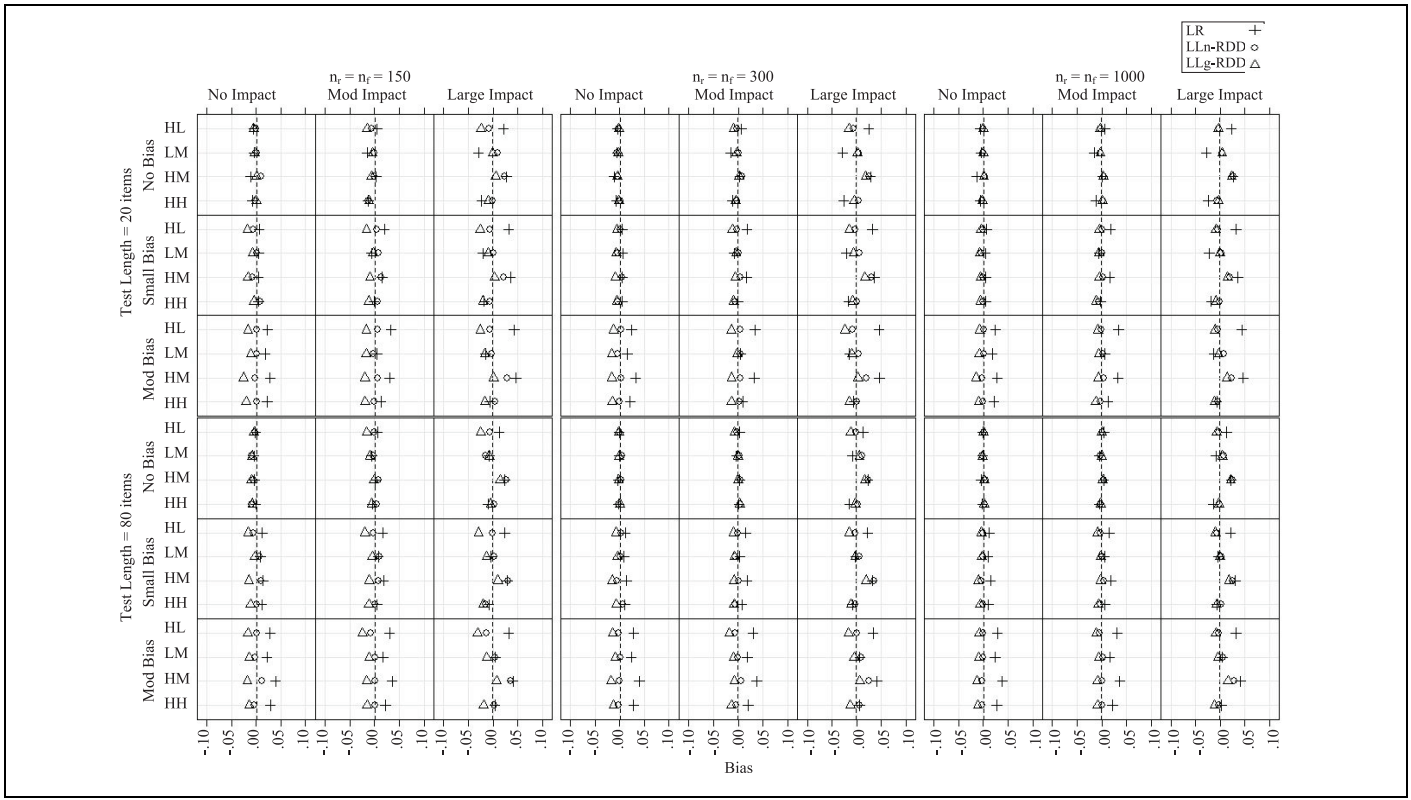

Figure 4. Monte Carlo Estimated Bias of p_{DIF} for DIF Items

Note. Dashed vertical line indicates where bias = 0. DIF = differential item functioning; LR = logistic regression; LLn-RDD = local linear regression, regression discontinuity design; LLg-RDD = local logit estimation, regression discontinuity design; n_r = sample size of reference group; n_f = sample size of focal group; n_r = sample size of reference group; n_f = sample size of focal group; Impact = latent mean group difference (0 *SD*, .5 *SD* [Mod], 1 *SD* [Large]); Bias = selection bias; HL = item with high discrimination (a = 1.6), low difficulty (b = −1.5); LM = item with low discrimination (a = 0.6), moderate difficulty (b = 0.0); HM = item with high discrimination (a = 1.6), moderate difficulty (b = 0.0); HH = item with high discrimination (a = 1.6), high difficulty (b = 1.5).

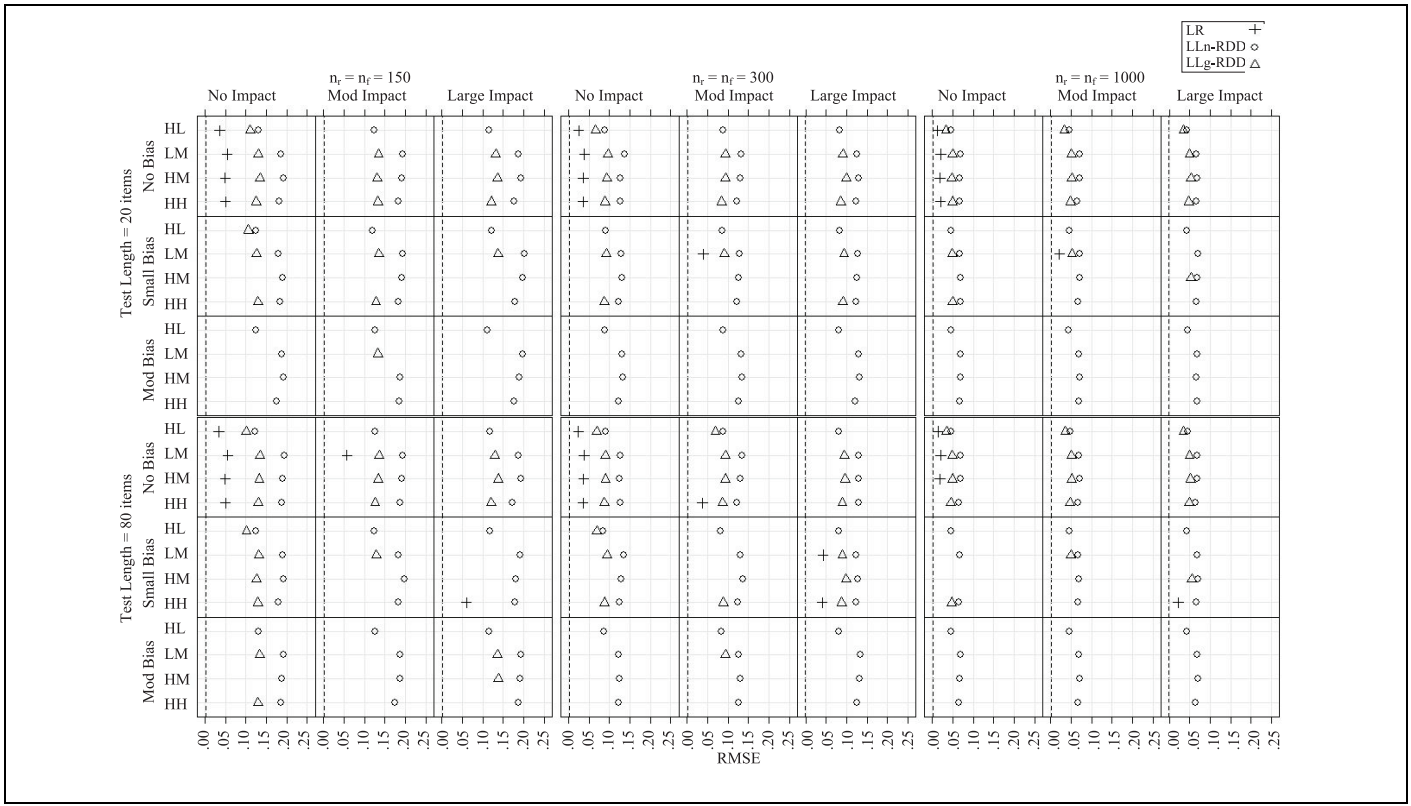

Figure 5. Monte Carlo Estimated RMSE of p_{DIF} for Non-DIF Items

Note. RMSE is only displayed where corresponding estimator is not significantly biased. Dashed vertical line indicates where RMSE = 0. RMSE = root mean square error; DIF = differential item functioning; LR = logistic regression. LLn-RDD = local linear regression, regression discontinuity design. LLg-RDD = local logit estimation, regression discontinuity design. n_r = sample size of reference group. n_f = sample size of focal group. n_r = sample size of reference group. n_f = sample size of focal group. Impact = latent mean group difference (0 *SD*, .5 *SD* [Mod], 1 *SD* [Large]). Bias = selection bias. HL = item with high discrimination (a = 1.6), low difficulty (b = −1.5). LM = item with low discrimination (a = 0.6), moderate difficulty (b = 0.0). HM = item with high discrimination (a = 1.6), moderate difficulty (b = 0.0). HH = item with high discrimination (a = 1.6), high difficulty (b = 1.5).

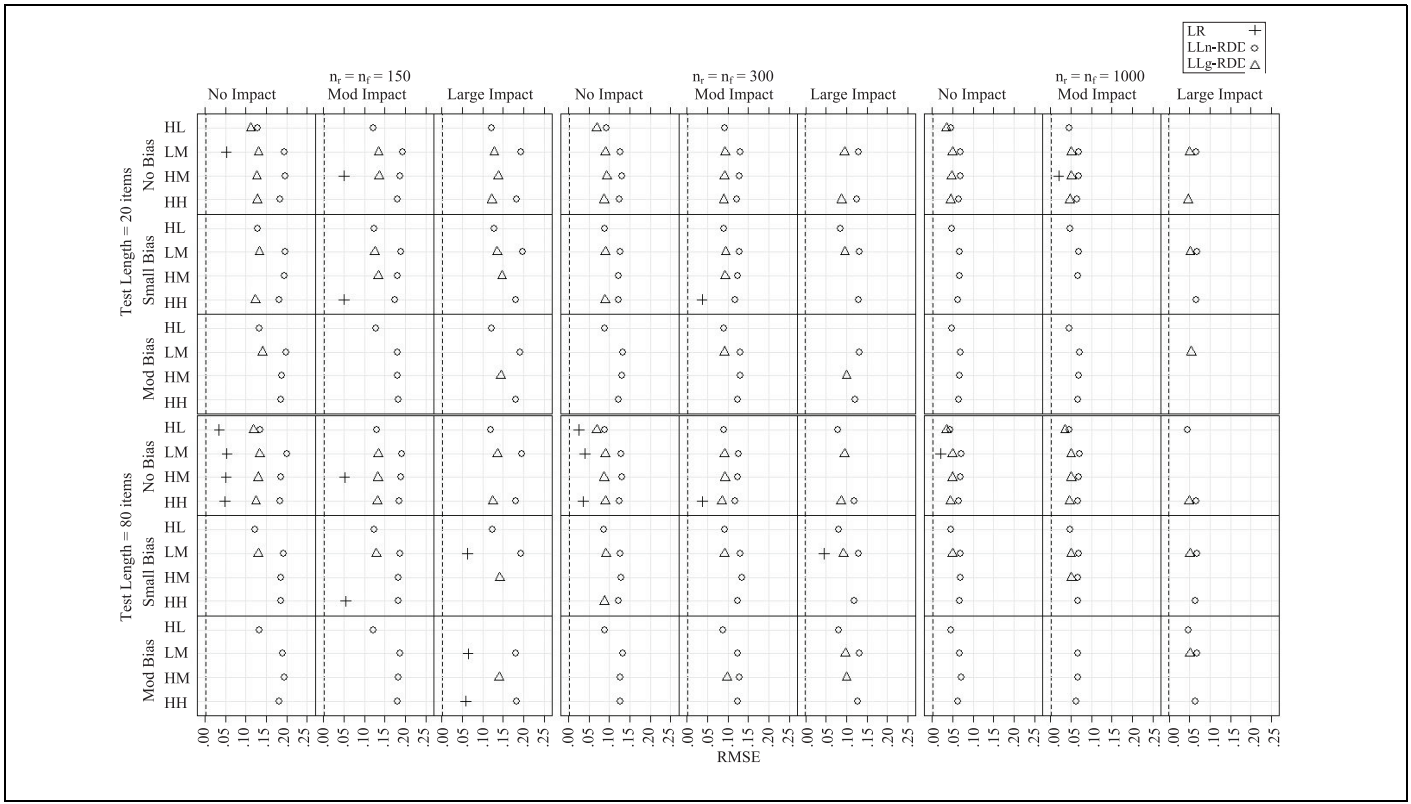

Figure 6. Monte Carlo Estimated RMSE of p_{DIF} for DIF Items

Note. RMSE is only displayed where corresponding estimator is not significantly biased. Dashed vertical line indicates where RMSE = 0. RMSE = root mean square error; DIF = differential item functioning; LR = logistic regression. LLn-RDD = local linear regression, regression discontinuity design. LLg-RDD = local logit estimation, regression discontinuity design. n_r = sample size of reference group. n_f = sample size of focal group. n_r = sample size of reference group. n_f = sample size of focal group. Impact = latent mean group difference (0 *SD*, .5 *SD* [Mod], 1 *SD* [Large]). Bias = selection bias. HL = item with high discrimination (a = 1.6), low difficulty (b = −1.5). LM = item with low discrimination (a = 0.6), moderate difficulty (b = 0.0). HM = item with high discrimination (a = 1.6), moderate difficulty (b = 0.0). HH = item with high discrimination (a = 1.6), high difficulty (b =1.5).

Discussion

Our objectives in this article were to develop and describe two approaches for evaluating DIF within an RDD framework and compare these novel approaches with traditional LR. We achieved our first objective by proposing the use of nonparametric local linear regression and local logit estimation within an RDD continuity-based framework (LLn-RDD and LLg-RDD, respectively) to evaluate uniform DIF. We achieved our second objective by performing a Monte Carlo simulation study that compared the Type I error and power rates of the LR, LLn-RDD, and LLg-RDD uniform DIF test statistics, and bias and RMSE of the LR, LLn-RDD, and LLg-RDD uniform DIF effect size estimators.

Comparison of LR, LLn-RDD, and LLg-RDD for Evaluating DIF

As hypothesized, the LLn-RDD and LLg-RDD uniform DIF test statistics had less inflated Type I error rates (never exceeding .09 and .06, respectively) than the corresponding LR test statistic (reaching as high as .93). The LLn-RDD and LLg-RDD statistics were relatively stable across conditions, although the LLg-RDD statistic was overly conservative (Type I error rate $< .03$) at times, consistent with Xu's (2017) findings that the local logit standard errors were inflated. In line with prior research the LR statistic was sensitive to selection bias (Liu et al., 2019), sample size (DeMars, 2009, 2010; Y. Li et al., 2012; Shih et al., 2014), and the strength of the association between the item and underlying latent trait (DeMars, 2010; Rogers & Swaminathan, 1993). The finding that the LR statistic was more sensitive to selection bias when the item was strongly discriminating is unsurprising based on Equation 4 in which the selection bias parameter is multiplied by the discrimination parameter. Assuming a testing context that mirrors our simulation study, if there is a moderate level of selection bias, sample size is large, and the item has a high item-ability correlation, the probability of flagging the item for uniform DIF, when in fact the item does not exhibit DIF, is greater than .50. Such a high false positive rate has serious implications for the test construction phase in which unnecessary time and money may be devoted to reviewing the flagged items, and well-functioning items that take time and money to develop and replace may be errantly thrown out.

Contrary to our hypothesis, group impact and test length did not account for a meaningful proportion of variability in Type I error rates. However, focusing on the conditions with no selection bias, the pattern of results shown in Figure 1 is consistent with prior research, indicating that LR Type I error rates are inflated when the matching score is unreliable (when the test is short) and group impact is large, particularly when sample size is large (DeMars, 2009, 2010).

As expected, considering only those conditions in which the Type I error rate of the uniform DIF test statistic did not exceed the 99% confidence bounds for a true Type I error rate of .05, the LR statistic was considerably more powerful than the corresponding LLn-RDD and LLg-RDD statistics (by .56 and .54, on average, respectively). LLn-RDD demonstrated slightly greater power than LLg-RDD, despite

smaller effective sample sizes. Even under the largest sample size condition, power of the LLn-RDD and LLg-RDD statistics to detect a moderate level of uniform DIF never reached .80 and was less than .30 for the two items with low item-ability correlations. While false positives are costly, failing to detect DIF when an item truly does function differently across groups (a false negative) is doubtlessly more problematic in educational contexts in which the end goal is to achieve unbiased and equitable testing. Consistent with prior research, across approaches power was highest when sample size was large and the item was strongly correlated with the underlying latent trait (e.g., Z. Li, 2014).

Consistent with our hypothesis, the LLn-RDD and LLg-RDD effect size estimators were less biased than the LR estimator in the presence of selection bias and when impact was large for the two items with high item-ability correlations. However, when considering the p metric classification system presented in Monahan et al. (2007) that distinguishes among $|p| \leq .05$, $.05 < |p| \leq .10$, and $|p| > .10$, the level of bias was relatively minor for all three approaches across most conditions. Bias was at or below .05 for 94% of the conditions under the LR approach and below .05 for all conditions under the LLn-RDD and LLg-RDD approaches. These results suggest that, in expectation, the estimated magnitude of p_{DIF} will not be far from the true value, even if inferences are untrustworthy under those same conditions. This reiterates the importance of considering both statistical significance and effect size when evaluating DIF.

Finally, only considering the conditions in which the effect size estimators were unbiased, the LR estimator was notably more precise than the LLn-RDD and LLg-RDD estimators (by .09–.11 and .05–.06 on average, respectively). RMSE of the LLn-RDD estimator averaged .12 across conditions and reached as high as .20 when sample size was small. That is, for any given sample, under these same conditions, the LLn-RDD estimated magnitude of p_{DIF} is expected to differ from the true value of p_{DIF} on average by as much as .20. These values are on the probability scale and thus represent considerable variability. Consistent with Frölich's (2006) findings, the LLg-RDD estimator was more precise than the LLn-RDD estimator. Unsurprisingly, across approaches, greater precision was observed when the sample size was large and the item was strongly correlated with the underlying latent trait.

Taken together, these results corroborate prior research demonstrating limitations of the LR DIF test statistic, specifically its high rate of false positives under certain conditions. Whereas the novel LLn-RDD and LLg-RDD DIF approaches posed theoretical advantages for addressing these limitations, they suffered from low statistical power and lack of precision.

Recommendations for Practice

Our first recommendation in choosing a DIF framework is to reflect on the testing context. Does the testing context lend itself to an RDD analysis (i.e., was group membership determined on the basis of a pretreatment running variable and preestablished

cut point)? If not, then LLn-RDD and LLg-RDD cannot be applied. As we note in the Introduction, this framework, in its current form, lends itself most directly to large-scale research testing contexts in which a single cut score is used to determine accommodations, in contrast to educational practice in which typically multiple sources of information are used. What is the research question—Is estimating the average treatment effect at the cut point even appropriate/desired? For example, if the aim is to test whether items function differently for students who are proficient versus not proficient in English, then evaluating DIF across language forms at the English language proficiency cut point is clearly inappropriate. What are the relative costs of a Type I versus Type II error? If Type I errors are not particularly costly, then LLn-RDD and LLg-RDD do not offer a distinct advantage. What are the testing conditions (e.g., sample size, item properties)? Group sizes need to exceed 1,000 to have sufficient power ($> .80$) to detect a moderate degree of uniform DIF. DIF items that are only weakly discriminating are unlikely to be flagged. (It could be argued, however, that weakly discriminating items are likely to be discarded early in the test construction process, making this point moot.)

In line with the advice of Hambleton (2006), our second recommendation is to use multiple approaches and multiple types of information (statistical significance, effect size) to evaluate DIF. LLn-RDD and LLg-RDD were found to have low power and precision, but may still be useful as a means for exploring the presence of, and sensitivity of inferences to, selection bias. RD plots provide a graphical depiction of (dis)continuity in outcomes or pretreatment covariates at the cut point by plotting the test-taker's score on the target variable (y-axis) in relation to the test-taker's value on the running variable (x-axis). A clear discontinuity in the probability of a correct response at the running variable cut point for an item under investigation for DIF suggests the presence of uniform DIF that can be attributed to the different test forms. On the contrary, a positive or negative association between the running variable and item response that is continuous (does not jump) at the running variable cut point suggests that inferences based on traditional approaches for evaluating DIF may be confounded by selection bias. Overall, LLn-RDD and LLg-RDD performed similarly, but the LLg-RDD effect size estimator was slightly more precise and thus we recommend its use over LLn-RDD for quantifying the magnitude of DIF.

Our third recommendation is that items flagged for DIF should be carefully reviewed by content experts, regardless of DIF approach. Although the proposed RDD framework supports causal inferences (e.g., that DIF is due to differences in the alternate language forms rather than differences in the test-takers assigned to the different forms), it does not provide an indication of the specific source of DIF (e.g., a problem with the translation of a particular word).

Limitations

Our simulation included many conditions, but certain factors were not considered that may influence the performance of the LR and RDD approaches. Most notably, we

did not consider nonuniform DIF. The RDD approaches are unlikely to be sensitive to nonuniform DIF, in contrast to the LR approach that can detect both types of DIF. In addition, we held constant the magnitude and direction of DIF, proportion of DIF items, and generating model, and we made the simplifying assumption of equal group sizes. The RDD approaches had low power and a lack of precision for detecting a moderate level of DIF under the ideal scenario of equal group sizes; they are expected to perform even worse for detecting smaller magnitudes of DIF and when group sizes are unequal. In contrast to large-scale research studies, more complex testing contexts in which multiple factors determine assignment to form are typical of educational practice and our simulation is not able to inform such contexts. We also generated the data so that all assumptions underlying the RDD approaches were met. In practice, these assumptions must be tested and are not always met. For example, it may be possible for test-takers or test administrators to manipulate scores on the running variable to influence group assignment. In this case, test-takers just below and above the cut point may not be similar on all pretreatment covariates. It is also possible, and indeed likely, that the running variable is measured with error.

Another limitation is that we considered only one type of effect size, the group difference in the predicted proportion of a correct response (p metric). While the p metric is easy to interpret, it is not constant across items with different difficulty levels and it is not a natural effect size estimator for the LR approach (in contrast to the conditional odds ratio).

Future Directions

In addition to evaluating other simulation conditions described in the ''Limitations'' section, our proposed framework for detecting DIF can be expanded and improved upon in multiple ways. It is particularly imperative to extend the framework to support investigations of nonuniform DIF and to improve power and precision. To this end, a parametric RDD approach may be considered, which would be comparable to the covariate approach for controlling selection bias that was described in the Introduction. Another possible extension is to generalize inferences about DIF beyond the running variable cut point (e.g., by utilizing multiple cut points). Other future directions include extending the framework to support multiple running variables and fuzzy RDDs in which the running variable cut point is not deterministic (Bloom, 2012) and using alternative rules to flag items for DIF that take into account both statistical significance and effect size (cf. Hidalgo et al., 2014; Jodoin & Gierl, 2001). Finally, other DIF frameworks for controlling selection bias, beyond RDD, should be considered.

Conclusion

The findings of our simulation study highlight the importance of considering selection bias when evaluating items for DIF. Due to low power and lack of precision, we

do not recommend relying exclusively on the newly proposed framework (at least not in its current form) when the testing context mirrors the conditions evaluated in our study. False negatives have significant implications for equity in educational assessment as failure to account for problematic items could result in the use of a test accommodation that unfairly advantages one group of students over another (e.g., if items displaying DIF are systematically easier for one group). However, we do advocate its use as an exploratory tool that can help evaluate the sensitivity of traditional methods for testing DIF, given clear evidence of selection bias in real-world testing scenarios in which alternate form assessment accommodations are used (see Goodrich et al., 2021). Additional methodological research is needed to improve the proposed framework.

Declaration of Conflicting Interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a grant from the American Educational Research Association (AERA), which receives funds for its ''AERA Grants Program'' from the National Science Foundation (NSF) under NSF award NSF-DRL No. 1749275. Opinions reflect those of the authors and do not necessarily reflect those of AERA or NSF.

ORCID iD

Natalie A. Koziol https://orcid.org/0000-0003-3275-1776

Supplemental Material

Supplemental material for this article is available online.

Note

1. For example, if the overall mean of the latent variable was held constant at 0 across unbalanced sample size conditions, then the group means would not be symmetric around 0 for the conditions with group differences in the latent variable. This asymmetry would lead to differences in item and test information across unbalanced sample size conditions.

References

Aikens, N., West, J., McKee, K., Moiduddin, E., Atkins-Burnett, S., & Xue, Y. (2020). Screening approaches for determining the language of assessment for dual language learners: Evidence from Head Start and a universal preschool initiative. *Early Childhood Research Quarterly*, *51*, 39–54. https://doi.org/10.1016/j.ecresq.2019.07.008

Bandel, E., Atkins-Burnett, S., Castro, D., Smither Wulsin, C., & Putnam, M. (2012, June). *Examining the use of language and literacy assessments with young dual language learners* [Report submitted to the University of North Carolina, FPG Child Development Institute, Center for Early Care and Education—Dual Language Learners]. Mathematica Policy Research.

Bloom, H. S. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness*, *5*(1), 43–82. https://doi.org/10.1080/19345747.2011.578707

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2021). *rdrobust: Robust data-driven statistical inference in regression-discontinuity designs* [Computer software] (R Package Version 1.0.7). https://CRAN.R-project.org/package=rdrobust

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2020a). *A practical introduction to regression discontinuity designs: Extensions*. Cambridge University Press.

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2020b). *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press.

Chen, M. Y., Liu, Y., & Zumbo, B. D. (2020). A propensity score method for investigating differential item functioning in performance assessment. *Educational and Psychological Measurement*, *80*(3), 476–498. https://doi.org/10.1177/0013164419878861

Clauser, B. E., Nungester, R. J., & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measurement*, *33*(4), 453–464. https://doi.org/10.1111/j.1745-3984.1996.tb00501.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

DeMars, C. E. (2009). Modification of the Mantel-Haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics*, *34*(2), 149–170. https://doi.org/10.3102/1076998607313923

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

Fidalgo, A. M., Alavi, S. M., & Amirian, S. M. R. (2014). Strategies for testing statistical and practical significance in detecting DIF with logistic regression models. *Language Testing*, *31*(4), 433–451. https://doi.org/10.1177/0265532214526748

Frölich, M. (2006). Non-parametric regression for binary dependent variables. *The Econometrics Journal*, *9*, 511–540. https://doi.org/10.1111/j.1368-423X.2006.00196.x

Gierl, M. J., Gotzmann, A., & Boughton, K. A. (2004). Performance of SIBTEST when the percentage of DIF items is large. *Applied Measurement in Education*, *17*(3), 241–264. https://doi.org/10.1207/s15324818ame1703_2

Goodrich, J. M., Koziol, N. A., & Yoon, H. (2021). Are translated mathematics items a valid accommodation for dual language learners? Evidence from ECLS-K. *Early Childhood Research Quarterly*, *57*, 89–101. https://doi.org/10.1016/j.ecresq.2021.06.001

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care*, *44*(11), S182–S188. https://doi.org/10.1097/01.mlr.0000245443.86671.c4

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. https://doi.org/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). Lawrence Erlbaum.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Lawrence Erlbaum.

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. https://doi.org/10.1207/S15324818AME1404_2

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(2), 281–355. https://doi.org/10.1257/jel.48.2.281

Lee, S. (2017). Detecting differential item functioning using the logistic regression procedure in small samples. *Applied Psychological Measurement*, *41*(1), 30–43. https://doi.org/10.1177/0146621616668015

Lhamon, C. E., & Gupta, V. (2015, January 7). *Dear Colleague Letter: English learner students and limited English proficient parents*. Office for Civil Rights, U.S. Department of Education, Civil Rights Division, U.S. Department of Justice.

Li, Y., Brooks, G. P., & Johanson, G. A. (2012). Item discrimination and Type I error in the detection of differential item functioning. *Educational and Psychological Measurement*, *72*(5), 847–861. https://doi.org/10.1177/0013164411432333

Li, Z. (2014). Power and sample size calculations for logistic regression tests for differential item functioning. *Journal of Educational Measurement*, *51*(4), 441–462. https://doi.org/10.1111/jedm.12058

Liu, Y., Kim, C., Wu, A. D., Gustafson, P., & Kroc, E. (2019). Investigating the performance of propensity score approaches for differential item functioning analysis. *Journal of Modern Applied Statistical Methods*, *18*(1), eP2744. https://doi.org/10.22237/jmasm/1556669280

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

Magis, D., & De Boeck, P. (2012). A robust outlier approach to prevent Type I error inflation in differential item functioning. *Educational and Psychological Measurement*, *72*(2), 291–311. https://doi.org/10.1177/0013164411416975

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of non-uniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement*, *54*(2), 284–291. https://doi.org/10.1177/0013164494054002003

Mellenberg, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, *7*(2), 105–108. https://doi.org/10.2307/1164960

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

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. https://doi.org/10.3102/1076998606298035

Muthén, L. K., & Muthén, B. O. (1998–2020). *Mplus user's guide* (Version 8.5).

Najarian, M., Snow, K., Lennon, J., & Kinsey, S. (2010). *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool-kindergarten 2007 psychometric report* (NCES 2010-

009). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Najarian, M., Tourangeau, K., Nord, C., & Wallner-Allen, K. (2018). *Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), kindergarten psychometric report* (NCES 2018-182). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

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

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

Petersen, M. A., Groenvold, M., Bjorner, J. B., Aaronson, N., Conroy, T., Cull, A., Fayers, P., Hjermstad, M., Sprangers, M., & Sullivan, M. (2003). Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. *Quality of Life Research*, *12*(4), 373–385. https://doi.org/10.1023/a:1023488915557

Pitoniak, M. J., Young, J. W., Martiniello, M., King, T. C., Buteux, A., & Ginsburgh, M. (2009). *Guidelines for the assessment of English-Language Learners* (ETS Office of Professional Standards Compliance's Fairness Series). Educational Testing Service. https://www.ets.org/s/about/pdf/ell_guidelines.pdf

R Core Team. (2019). *R: A language and environment for statistical computing* [Computer software] (Version 3.6.1). R Foundation for Statistical Computing. https://www.R-project.org/

Rock, D. A., & Pollack, J. M. (2002). *Early Childhood Longitudinal Study-Kindergarten Class of 1998-99 (ECLS-K), psychometric report for kindergarten through first grade* (NCES 2002-05). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

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. https://doi.org/10.1177/014662169301700201

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Salvia, J., Ysseldyke, J., & Witmer, S. (2017). *Assessment in special and inclusive education*. Cengage Learning.

Shadish, W. R., & Steiner, P. M. (2010). A primer on propensity score analysis. *Newborn and Infant Nursing Reviews*, *10*(1), 19–26. https://doi.org/10.1053/j.nainr.2009.12.010

Shealy, R. T., & Stout, W. F. (1993). A model-biased 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. https://doi.org/10.1007/BF02294572

Shih, C.-L., Liu, T.-H., & Wang, W.-C. (2014). Controlling type I error rates in assessing DIF for logistic regression method combined with SIBTEST regression correction procedure and DIF-free-then-DIF strategy. *Educational and Psychological Measurement*, *74*(6), 1018–1048. https://doi.org/10.1177/0013164413520545

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

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, *51*(6), 309–317. https://doi.org/10.1037/h0044319

U.S. Department of Education. (2007). Title I: Improving the academic achievement of the disadvantaged: Individuals with Disabilities Act (IDEA); Final rule. *Federal Register*, *72*(67), 17747–17781.

U.S. Department of Education. (2021). *Early Childhood Longitudinal Studies (ECLS) program: Instruments & assessments*. https://nces.ed.gov/ecls/instruments2024.asp

Wu, A. D., Liu, Y., Stone, J. E., Zou, D., & Zumbo, B. D. (2017). Is difference in measurement outcome between groups differential responding, bias or disparity? A methodology for detecting bias and impact from an attributional stance. *Frontiers in Education*, *2*, Article 39. https://doi.org/10.3389/feduc.2017.00039

Xu, K.-L. (2017). Regression discontinuity with categorical outcomes. *Journal of Econometrics*, *201*(1), 1–18. https://doi.org/10.1016/j.jeconom.2017.07.004

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Lawrence Erlbaum.

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*. Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (Research Report ETS RR-12-08; pp. 1–30). Educational Testing Service.