EdWorkingPaper No. 25-1303

Beg to DIFfer: Resolving Statistical Complications of Intersectional DIF Analyses

Lily An
Georgia State University

Edward J Kim
Bentley University

Modern test developers conduct differential item functioning (DIF) analyses to ensure fairness in educational and psychological testing. To address previously unrecognized biases, researchers have recently demonstrated the importance of conducting intersectional DIF analyses that attend to the intersectional nature of test-takers' multiple identities. However, these intersectional DIF approaches overlook how overlapping identity categories affect the statistical validity of DIF analyses. As the related tests violate independence, typical p-value corrections used in intersectional DIF analyses such as Bonferroni yield overly conservative family wise error rates (FWER) which limit statistical power to identify true DIF. Additionally, DIF on one dimension can spuriously cause DIF to appear while testing another demographic dimension with high overlap, a phenomenon we call signal interference. These concerns are particularly aggravated in intersectional DIF. We offer an approach utilizing parametric bootstrapping that adjusts significance levels of DIF detection processes to yield the intended Type I error rates. Using simulations studies, we illustrate the statistical complications of intersectional DIF analyses and the ability of our proposed method to resolve them.

VERSION: October 2025

Suggested citation: An, Lily, and Edward J Kim. (2025). Beg to DIFfer: Resolving Statistical Complications of Intersectional DIF Analyses. (EdWorkingPaper: 25-1303). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/8wc6-cr34

Title

Beg to DIFfer: Resolving Statistical Complications of Intersectional DIF Analyses

Abstract

Modern test developers conduct differential item functioning (DIF) analyses to ensure fairness in educational and psychological testing. To address previously unrecognized biases, researchers have recently demonstrated the importance of conducting intersectional DIF analyses that attend to the intersectional nature of test-takers' multiple identities. However, these intersectional DIF approaches overlook how overlapping identity categories affect the statistical validity of DIF analyses. As the related tests violate independence, typical p-value corrections used in intersectional DIF analyses such as Bonferroni yield overly conservative family wise error rates (FWER) which limit statistical power to identify true DIF. Additionally, DIF on one dimension can spuriously cause DIF to appear while testing another demographic dimension with high overlap, a phenomenon we call signal interference. These concerns are particularly aggravated in intersectional DIF. We offer an approach utilizing parametric bootstrapping that adjusts significance levels of DIF detection processes to yield the intended Type I error rates. Using simulations studies, we illustrate the statistical complications of intersectional DIF analyses and the ability of our proposed method to resolve them.

Keywords:

differential item functioning, simulations, parametric bootstrapping,

intersectionality

1. Introduction

Conducting differential item functioning (DIF) analyses is best practice for test developers to promote fairness by flagging potentially biased items (AERA, APA, & NCME, 2014). DIF represents a statistically significant difference in test item performance for similar overall ability test-takers of different demographic groups, and test items which show DIF are sent for further review and possible removal. The comparison of interest is traditionally between a reference group of test-takers and a focal group of test-takers. Importantly, DIF supports valid test design by removing items that demonstrate possible construct-irrelevant variance due to group status, thereby mitigating item bias.

Recent scholarly work has sought to integrate an intersectional perspective into traditional DIF methods (Russell et al. 2021; Albano et al., 2024). Intersectionality, a term coined by Kimberlé Crenshaw (1991), refers to the multiplicity of identities that individuals each hold. While traditional DIF compares one sub-category of identity to another sub-category, such as the comparison of high-income to low-income test-takers, intersectional DIF analyses consider the interaction of identity categories, such as the comparison of high-income White test-takers to high-income Black, low-income White, and low-income Black test-takers (Russell & Kaplan, 2021). Intersectional DIF represents another avenue through which educational measurement researchers can conduct quantitative research that engages with fairness (Sireci, 2020) and aims to address systemic bias (Else-Quest & Hyde, 2016; Russell, 2023). By accounting for the interaction of test-takers' identities, intersectional DIF research can reveal instances of DIF that were unexplored in traditional DIF analysis. For example, Albano et al. (2024) find that traditional DIF, which they term *main effects* DIF, under-counts items with DIF as identified through intersectional DIF analysis; a conclusion arrived at separately by Russell et al. (2022). Thus, standard DIF practice may offer misleading evidence of measurement invariance, as some validated test items could in fact present bias against certain intersectional groups of test-takers.

One key consideration when conducting intersectional DIF analyses is how many comparisons between intersectional identities are made. Continuing the prior example, there were four comparisons made between high- and low-income White and Black test-takers. The total number of comparisons rises dramatically as the demographic categories under investigation, and the subcategories within identity categories, for DIF increase. Even when conducting traditional DIF, analysts need to be aware of multiple testing concerns – namely, that increasing the number of statistical tests conducted will increase the likelihood of encountering Type I error, i.e. that items will be flagged for DIF by chance. Though Type I error adjustments exist, such as the Bonferroni correction (Bonferroni, 1936) or the Benjamini-Hochberg false discovery rate (Benjamini & Hochberg, 1995), such methods require tests to be independent – or at least to calculate and account for the extent of dependence – to produce valid adjusted p-values. DIF detection has an additional multiple testing complication that some approaches do not benchmark to p-values. In fact, the first studies conducting intersectional DIF, Russell & Kaplan (2021) and Russell et al. (2021), specifically note not applying p-value corrections to their results because their DIF detection approaches were incompatible. Later intersectional DIF research has used methods that adjust for Type I error (Albano et al., 2024).

As the field of psychometrics continues to integrate intersectional perspectives in DIF analyses, two underdiscussed complications related to the dependence of tests become more problematic. These complications stem from the reuse of observations across tests in a typical battery of DIF analyses, as each demographic dimension is considered in turn. This leads to: 1) “overcorrection”, where misapplied multiple testing corrections will impose too severe a benchmark for significance, and 2) “signal interference”, where true DIF on one axis induces false identification of DIF on another. These complications function in opposite directions, yet both negatively affect the DIF detection process: overcorrection leads to less sensitivity to detect true DIF-containing items, and signal interference leads to spurious cases of DIF detection.

Using simulations, we first demonstrate the threat of each issue, illustrating how the extent of the problem depends on the distribution of observations across identities. We then introduce parametric bootstrapping, a tool used in many fields, as a way to flexibly model the interdependence of DIF tests and subsequently produce correctly adjusted benchmarks for significance while considering either issue. This research builds on existing work in the exploration of intersectional DIF and highlights technical concerns with using existing Type I error control methods. By raising and addressing key challenges in intersectional DIF, we fill a growing gap in the academic literature seeking to integrate psychometrics with intersectionality.

The paper is organized as follows: In Section 2, we provide literature on the methodology of DIF analysis as well as the use of p-value corrections in both traditional and intersectional DIF. Next, in Section 3 we detail our simulations parameters and explain our implementation of parametric bootstrapping in the present context. The Results section, Section 4, considers the overcorrection and signal interference issues in turn, demonstrating the extent of each issue and comparing the performance of parametric bootstrapping to typical correction methods. Sections 5 and 6 conclude.

2. Background

Detecting DIF has critical implications for the validity, fairness, and interpretability of test scores. Items exhibiting statistically significant DIF suggest that individuals from different groups who are matched on the underlying trait or ability do not have an equal probability of answering the item correctly. Flagged items are reviewed by content experts to determine whether the differential performance reflects construct-irrelevant variance (e.g., unfair bias) or construct-relevant differences (e.g., true group differences in the trait being measured).

Traditional DIF analyses compare the performance of a reference group of test-takers to a focal group of test-takers. For example, the reference group of high income test-takers may be compared to the focal group of low income test-takers. There are multiple approaches to estimate DIF, which can primarily be categorized into IRT-based and non-IRT based methods. We focus here on the widely used Mantel-Haenszel (MH) test, which is a non-IRT based method (Mantel & Haenszel, 1959; Holland & Thayer, 1986). The MH test assumes that the odds of getting an item correct across the distribution of ability scores θ are the same in both the focal and reference groups - therefore testing the hypothesis of a common odds ratio between groups. Because the MH procedure conditions on total score and assumes a constant odds ratio, the MH test detects uniform DIF, which occurs when one group has a consistent advantage in answering an item correctly across all levels of ability θ. In contrast, nonuniform DIF arises when the direction or magnitude of the advantage changes across ability levels, and detecting nonuniform DIF typically requires more flexible methods such as logistic regression or IRT-based modeling approaches.

Recent work in DIF analysis has expanded beyond single-variable group comparisons to accommodate considerations of intersectional DIF, which examines how combinations of demographic characteristics (e.g., gender × race × socioeconomic status) may interact to locate DIF. Traditional DIF methods that analyze group membership along one dimension of identity at a time may fail to detect items that function differently for specific subgroups at the intersection of multiple identities (Aryadoust et al., 2024). Russell and colleagues (e.g., Russell & Kaplan, 2021; Russell et al., 2021; Russell et al., 2022) demonstrate that intersectional DIF analyses detect a greater number of flagged items compared to traditional approaches, a finding that holds even while using the same DIF criteria and attempting to adjust for multiple testing. Their findings suggest that relying solely on main effects DIF risks missing DIF that emerges through the interaction of group variables.

2.1. P-value adjustments in DIF research

When multiple statistical tests are conducted, the likelihood of encountering a false positive, or Type I error, grows. DIF analyses, which test each item on an assessment, are at particular risk of at least one Type I error (Shaffer, 1995). Incorrectly identifying non-DIF items as containing DIF burdens the test development process, requiring additional resources to evaluate the superfluously flagged items, substitute with alternative items, or decrease the exam's statistical power if suitable replacements are unavailable.

One response is to set the standard for significance, not at the individual test level, but at the aggregate level. The probability of at least one Type I error across a set of tests, also known as the family-wise error rate, can be calculated as:

$$FWER = 1 - (1 - \alpha)^m$$

when considering m independent tests, each with a significance level of α. To set the FWER to the desired level, various methods can be employed. For instance, the commonly used Bonferroni correction adjusts the individual test's significance level to $\alpha_{adjusted} = \frac{\alpha}{m}$, and the resulting FWER approximately equals α. The false discovery rate, in contrast, estimates the proportion of false positives among all findings declared significant. Though studies (e.g., Stark et al., 2006) have shown that p-value corrections (e.g., Bonferroni) can effectively reduce the Type I error rate of DIF tests (e.g. MH), this caution comes at the expense of reduced power. Items that on their own show small-to-medium DIF amounts will go undetected by the MH test once adjusted by the Bonferroni correction (Penfield, 2001; Kim & Oshima, 2012). Insofar as statistical analyses must balance the threats of Type I against Type II error, employing an overly strict standard for significance not only misaligns the results from the intended tolerance, but also requires a larger sample size than necessary to achieve the desired outcomes.

Applying an intersectional perspective to DIF analyses will typically generate more comparisons than traditional DIF, in turn yielding a greater likelihood of Type I errors. Thus far, intersectional DIF investigations that engaged with the multiple testing issue have employed the Bonferroni correction (Russell, Szendey, & Li, 2022; Albano et al., 2024). However, as we explain in subsequent sections, any p-value correction method that assumes independence across tests, or rigidly quantifies the extent of dependence, will be incorrect.

2.2. *Statistical complications in intersectional DIF analyses*

We highlight two issues with conducting DIF investigations in the intersectional context. The first concern involves the *overcorrection* of p-value adjustments typically used to control the Family Wise Error Rate (FWER) in DIF analyses. A core vulnerability of intersectional DIF is the considerably greater number of tests employed. However, if overlapping subsets of examinees are used across different tests, the tests violate independence, and adjusted p-values from methods such as Bonferroni can become overly conservative, exacerbating an already difficult standard for significance. Methods for p-value adjustment that allow the user to quantify the level of dependence across tests offer little help as the dependencies between DIF tests elude a simple closed form summary, conditional on not only demographic overlap but also the distribution of ability across demographic dimensions.

The second complication arises when there exists a true DIF effect. Suppose that an exam item engenders DIF across the income axis, and there exists a strong association between overall test scores for test-takers who hold certain race and income identities. With sufficient power, we would expect DIF tests to flag said problematic item as containing DIF on income as an axis of interest. But given the high overlap between income and race within the tested sample, we would be unsurprised if DIF tests also flagged race as an axis of interest, insofar as the sets of examinees used to compare performance across income groups largely resembled the sets used to compare performance across race categories. Most DIF tests cannot consider multiple demographic axes simultaneously and therefore cannot adjust expectations when signal from one is likely to interfere when detecting signal on another.

2.3. *Other DIF approaches for multiple identities fail to address p-value correction*

concerns

Other DIF approaches exist that can consider multiple identities, though their relevance depends on context. First, multidimensional IRT (MIRT) can be used to control for latent group-specific traits and check whether items behave differently across groups after accounting for multidimensionality. As Ackerman & Ma note, interpretation of flagged items is difficult (2024), and MIRT may be more useful when DIF is driven by multidimensionality as opposed to bias. Second, multiple indicators – multiple causes (MIMIC) models are structural equation models that detect DIF as direct effects from observed covariates to individual items. These measurement models are useful for exploratory analyses of group effects, as they only model mean impact, assume constant variance of θ, and do not allow for the distribution of the latent trait to depend on background variables (Bauer, 2023). Moderated nonlinear factor analysis (MNLFA) is a highly generalized method for detecting DIF, able to model multiple background variables simultaneously (Bauer, 2017; Bauer, 2020). While MNLFA could provide a broader view of DIF and its sources, at present MNFLA is less interpretable for assessment developers than commonly-used DIF methods like MH, logistic regression, Lord's chi-squared, and SIBTEST, to name a few. There are also stronger sample size requirements to use MNFLA in its fully unconstrained specification, and sequentially checking more sparse specifications merely relocates the issue of multiple hypothesis testing.

2.4. Parametric bootstrapping

We propose a method that uses parametric bootstrapping to adjust significance levels. Parametric bootstrapping is a resampling-based method used to approximate the sampling distribution of a statistic by generating repeated samples from a fitted parametric model. In this approach, data are simulated repeatedly from a fitted IRT model (Equation 1) under the null hypothesis of no DIF, using the estimated item and ability parameters. Each simulated dataset is then analyzed using the same DIF detection method, and the distribution of the test statistic across replications is used to estimate empirical p-values. Parametric bootstrapping has been used for decades (e.g., Self & Liang, 1987) as a way to help guard against Type I error inflation and misestimation of standard errors.

Recent work in DIF detection has begun to leverage parametric bootstrapping to improve the accuracy of test statistics and confidence intervals under nonstandard conditions. For example, a method within a MIMIC + 2PL framework that does not require prior anchor items retrieves item-level p-values and confidence intervals through a bootstrap procedure. (Chen et al., 2023). Software tools also facilitate parametric bootstrap in IRT: the *mstDIF* package provides the *bootstrap_sctest* function to compute score tests for DIF across item difficulty and discrimination parameters under covariates generated via parametric bootstrap (Debelak & Debeer, 2024); similarly, *mirt* offers a *boot.LR* function for bootstrap likelihood ratio tests between nested IRT models (Chalmers, 2012). We apply this tool to the intersectional DIF context to address the complications of overcorrection and signal interference.

3. Study Design

3.1. Data generation: demographics and item responses

Our simulation's data generating process simulates a context where a group of 1,000 examinees completes a 20-item exam. Using R, we generate examinees and randomly assign examinee demographics across two dimensions: a binary "income" variable (low-income and high-income), and a trinary "race/ethnicity" variable (Black, White, and Other). We use these group labels to reflect common groupings of people in social science contexts. We manipulate the demographic proportions of test-takers across the two dimensions by adjusting the number of "White high-income" versus "Black high-income" examinees, ranging from almost 0 and almost 1/3 respectively, to the opposite. The other four identities are fixed as close to 1/6 as possible; remainders are randomly assigned. We next generate dichotomously scored test data with a two parameter IRT model (see Equation 1 below). Examinee ability parameters are drawn from a standard Normal distribution, item difficulty parameters are drawn from a uniform distribution between -2 and 2, and item discrimination parameters are fixed at 1.

Following classic IRT, we model the probability of a correct response for an examinee *i* and item *j* via the following:

$$P\left(Y_{ij} = 1|\theta_j\right) = \frac{\exp\left(\mathrm{a_j}\left(\theta_j - b_i\right)\right)}{1+\exp\left(\mathrm{a_i}\left(\theta_j - b_i\right)\right)} \qquad \text{(Equation 1)}$$

3.2. DIF level and detection

To demonstrate the "overcorrection" issue, the data generating process contains no actual DIF. While demonstrating the "signal interference" issue, the simulation includes uniform DIF of 0.3 magnitude in favor of high-income examinees on item 1 only. We include DIF following Equation 2:

$$P\left(Y_{ij} = 1\middle|\theta_j^*\right) = \frac{\exp\left(a_j\left(\theta_j^* - b_i\right)\right)}{1 + \exp\left(a_j\left(\theta_j^* - b_i\right)\right)} \text{ where } \theta_j^* = \theta_j + \beta_I DIF_{1,j}^I + \beta_R DIF_{1,j}^R + \beta_X DIF_{1,j}^X$$

(Equation 2)

Where I represents income-based DIF between White and Black examinees, R represents race-based DIF between high-income and low-income examinees, and X represents intersectional DIF between White high-income and Black low-income examinees. We therefore test, in turn, the hypotheses $H_0\colon \beta_I = 0, H_0\colon \beta_R = 0, and\ H_0\colon \beta_X = 0$ for overcorrection, and the hypotheses $H_0\colon \beta_R = 0, and\ H_0\colon \beta_X = 0$ for signal interference. Our method of DIF detection is the widely employed Mantel-Haenszel test. Note that though the MH test's "continuity correction" is a popular option in practice, for clarity we do not use it as this option is known to distort Type I errors, the focus of the present study.

Our choice of comparisons was made for the sake of readability, though it is admittedly unrealistic in practice. In our context there exist 19 possible pairwise comparisons across main and interacted identities: traditional DIF would only consider some main effects, and a strictly intersectional DIF perspective would only consider the intersectional effects. However, the logic behind the statistical complications we expound in this study extends to any set of DIF tests, and this chosen set of three comparisons fully expresses the implications of our investigation while minimizing variance and illegibility. We address this point in greater detail in a later section.

3.3. Parametric bootstrapping and comparison

Though we use the MH test and the resulting MH statistic and p-value to flag items for DIF, we fit Equation 1, a logistic regression model, to estimate the parametric bootstrapping parameters (i.e. examinee ability, and item 1 difficulty and discrimination). We then repeat the earlier data generation process, now employing the estimated values to randomly generate new responses to item 1, while preserving for each examinee their demographic information and responses to items 2 through 20. Note that while demonstrating signal interference, we augment Equation 1 to include a term for DIF on income. For each bootstrapped data set, we run the same battery of MH tests and collect the results. In this way, the bootstrapped data sets exactly delineate the probability space on item 1 given the structure of the data set, and testing for DIF across bootstrapped data sets generates a distribution of MH test results expected under the chosen null hypothesis.

We calculate corrected p-values by comparing the originally observed estimate of interest against the bootstrapped distribution. For example, the adjusted p-value for the FWER is calculated by first collecting, across all bootstrapped data sets *m*, the lowest p-value across the three MH tests, p*_m, and then finding the proportion of p*_m that are less than each of the three observed p-values. Therefore, by construction, at least one of the three observed p-values will be less than five percent of the p*_m approximately five percent of the time under the null hypothesis of no DIF, which matches the definition of the desired FWER.

To calculate adjusted p-values for signal interference, we compare the observed MH statistic from each test against the corresponding MH statistics across the bootstrapped data sets, where the bootstrapped data sets also induce DIF on income to the extent estimated by the logistic regression model. Thus, the MH statistics generated from the bootstrapped data sets describe what one should typically expect from testing dimensions other than income, given DIF on income. If an observed MH statistic exceeds the 95th percentile of the generated statistics, then it has demonstrated a level of DIF that only would occur only five percent by chance under the null hypothesis of DIF on income to the extent estimated but no DIF on any other dimension.

4. Results

4.1. Overcorrection

We first demonstrate the fact that Bonferroni adjustments can be overly conservative for DIF identification in Figure 1 below. The simulated data contains no actual DIF, yet we conduct a pair of DIF tests that check across the race and income identity groups: Black versus White and low-income versus high-income tests. Two independent tests with a critical value of 0.05 would both show significance 0.05*0.05 = 0.0025 proportion of the time on average. Not only does the red line in Figure 1 show evidence of dependence between the race and income DIF tests by being off the dashed 0.0025 line, but the extent to which the tests agree differs across demographic disproportion along the x-axis. Namely, Figure 1 shows greater conflation of tested subgroups towards the right-hand side of the x-axis, where more of the high-income examinees are White, leading to greater agreement between tests' conclusions. We additionally include parallel results for the pair Black versus White and Black low-income versus White high-income (the green "Race & Intersection" line), as well as the pair low-income versus high-income and Black low-income versus White high-income (the blue "Income & Intersection" line). These lines, which sit well above 0.0025, show that spurious identifications of DIF can be dependent across tests, and even more so, with intersectional DIF analyses.

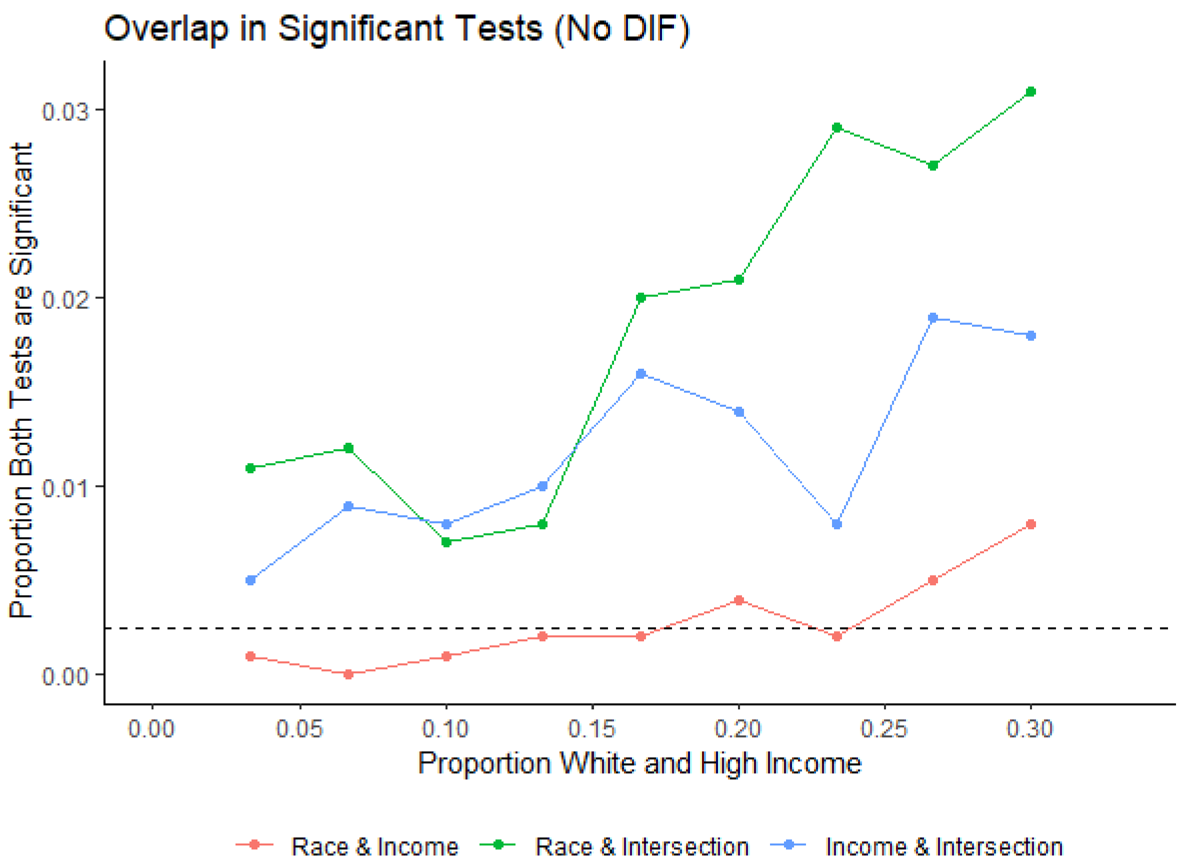

Figure 1. Rates of spurious DIF detection between pairs of DIF tests when identities are conflated within examinees. Dashed line marks 0.0025, the expected rate that two independent tests would both flag DIF.

This has implications for the Family Wise Error Rate (FWER), i.e. whether any of the tests meet significance, which practitioners often benchmark to the originally intended critical value. As shown in Figure 2 below, without any p-value correction in the blue line results, the FWER is understandably above 0.05, insofar as the chance of any one of multiple tests reaching significance is higher than the chance of one test reaching significance. But using the Bonferroni correction over-adjusts this rate, based as it is on the approximation that, under independence, the chance that at least one test reaches significance by chance under the null hypothesis is $1 - (0.95)^{m} \approx 0.05m$.

Figure 2 validates the proposed method, showing that approximately five percent of trials showed significance when benchmarked to the critical value resulting from parametric bootstrapping. Bonferroni, which we hypothesized as overly conservative, indeed systematically falls below the intended FWER (shown by the dotted line), indicating that that threshold for significance is too strict and the balance between Type I and Type II error rates is misaligned from expectation.

Importantly, this pattern holds across demographic disproportion. Regardless of how much sample overlap exists between the three MH tests, the parametric bootstrap method correctly adjusts expectations relative to the distribution of demographics and ability in the given data. On the other hand, we notice that the Bonferroni correction appears increasingly misaligned as the overlap between the White racial category and the high-income category grows towards the right-hand side of the x-axis in Figure 2. Greater sample overlap implies greater deviation from three independent tests, and thus the Bonferroni method to divide by that number of tests becomes less appropriate as the MH tests results grow more redundant with regard to each other.

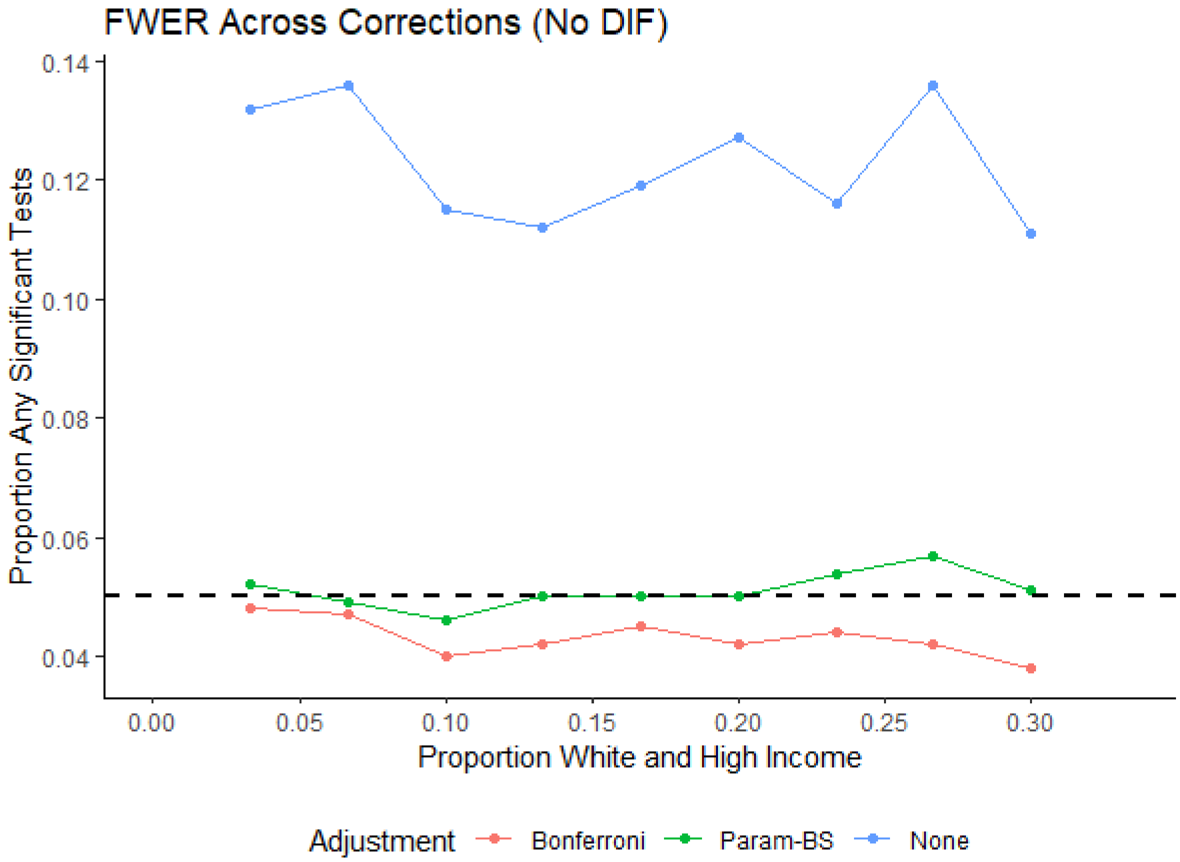

Figure 2. Rates of any DIF being detected using no p-value correction, Bonferroni correction, and parametric bootstrapping. Dashed line represents the familywise error rate (FWER).

4.2. Signal Interference

Under our simulations, item 1 exhibits true income-based DIF in favor of high-income examinees. There is no direct DIF from sources of racial identity or in the examined intersectional identities. Figure 3 illustrates that the false discovery rate for the DIF-free race and race-by-income comparisons is not only misaligned with the alpha value of 0.05, but also dependent upon the extent of demographic disproportion. Regardless of p-value correction approach, comparisons across race (the solid lines) are more likely to indicate DIF at the extremes of demographic disproportion, as indicated by upticks in the lines at either end. On the right-hand side, the high-income identity is dominated by test-takers assigned to the White racial category, meaning that comparisons of White versus Black test-takers overlap with comparisons of high-income versus low-income test-takers in favor of White examinees. On the left-hand side, the opposite is true, where the high-income identity is dominated by the Black racial category, and thus comparing across race flags DIF in favor of Black examinees. Similarly, the intersectional comparison, White high-income versus Black low-income, demonstrates greater likelihood of significance as the overlap between the White racial category and the high-income category increases, a logical consequence of the intersectional comparison having greater resemblance to the strictly-income comparison.

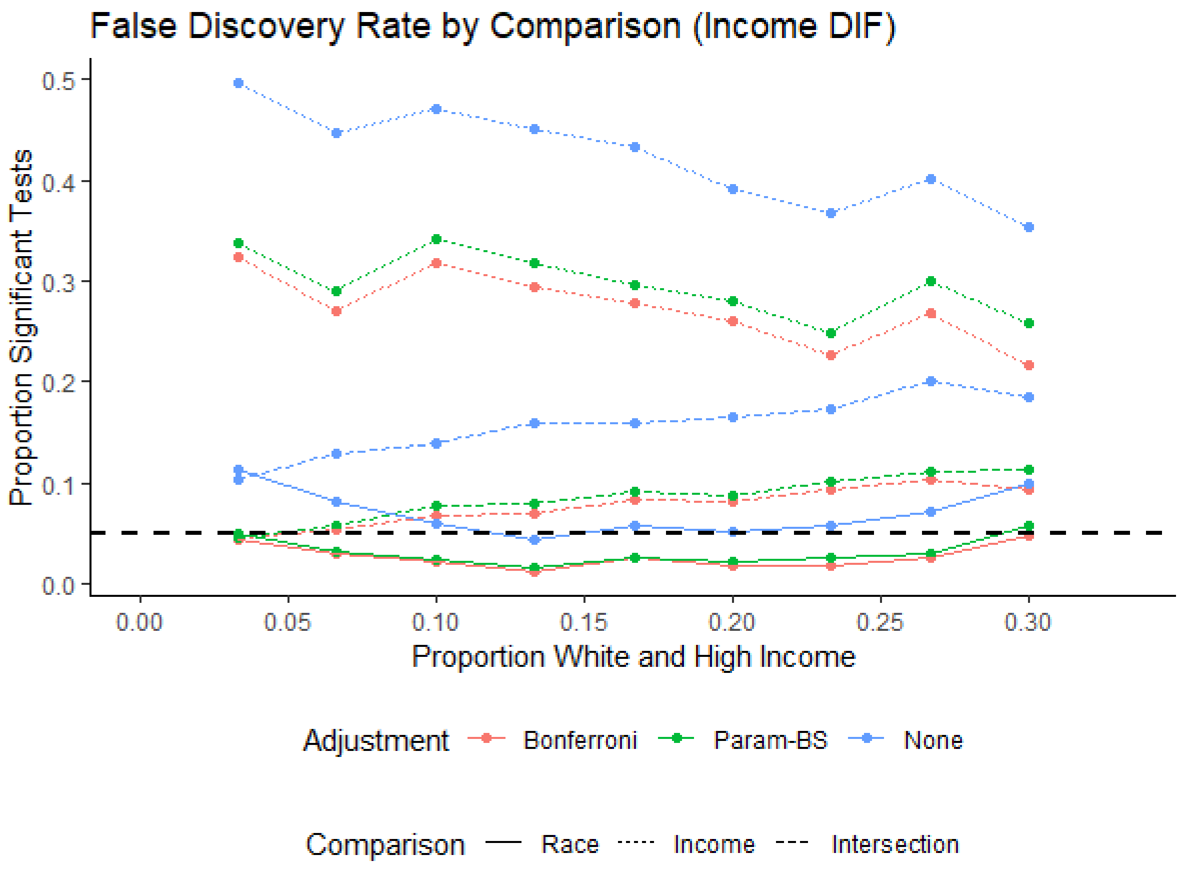

Figure 3. Rates of DIF identifications on race, income, and their intersection using no p-value correction, Bonferroni correction, and parametric bootstrapping, with DIF only truly existing on income. Dashed line represents the critical value.

Most DIF tests cannot control for multiple demographic dimensions simultaneously and therefore cannot adjust expectations when signal from one is likely to interfere with detecting signal of another. Our previous approach to parametric bootstrapping successfully adjusted benchmarks for the FWER under the null hypothesis because we generated the bootstrapped data sets under the null hypothesis. If we wish to adjust benchmarks when one effect is significant, we must generate bootstrapped data sets under this new condition.

Figure 4 illustrates that our proposed method for handling signal interference, also parametric bootstrapping, properly adjusts the false discovery rate for race and intersectional comparisons in the face of DIF on income. In this figure, we show the result of counting the instances of identified race-based and intersectional DIF, though DIF was only induced along the income dimension. Note that not only is the error rate at approximately five percent for tests, but this result seems largely independent of the amount of demographic disproportion per the green solid and dashed lines, proving that parametric bootstrapping flexibly adjusts to the exact context. In comparison, the Bonferroni correction is highly dependent on the distribution of demographic overlap across the x-axis.

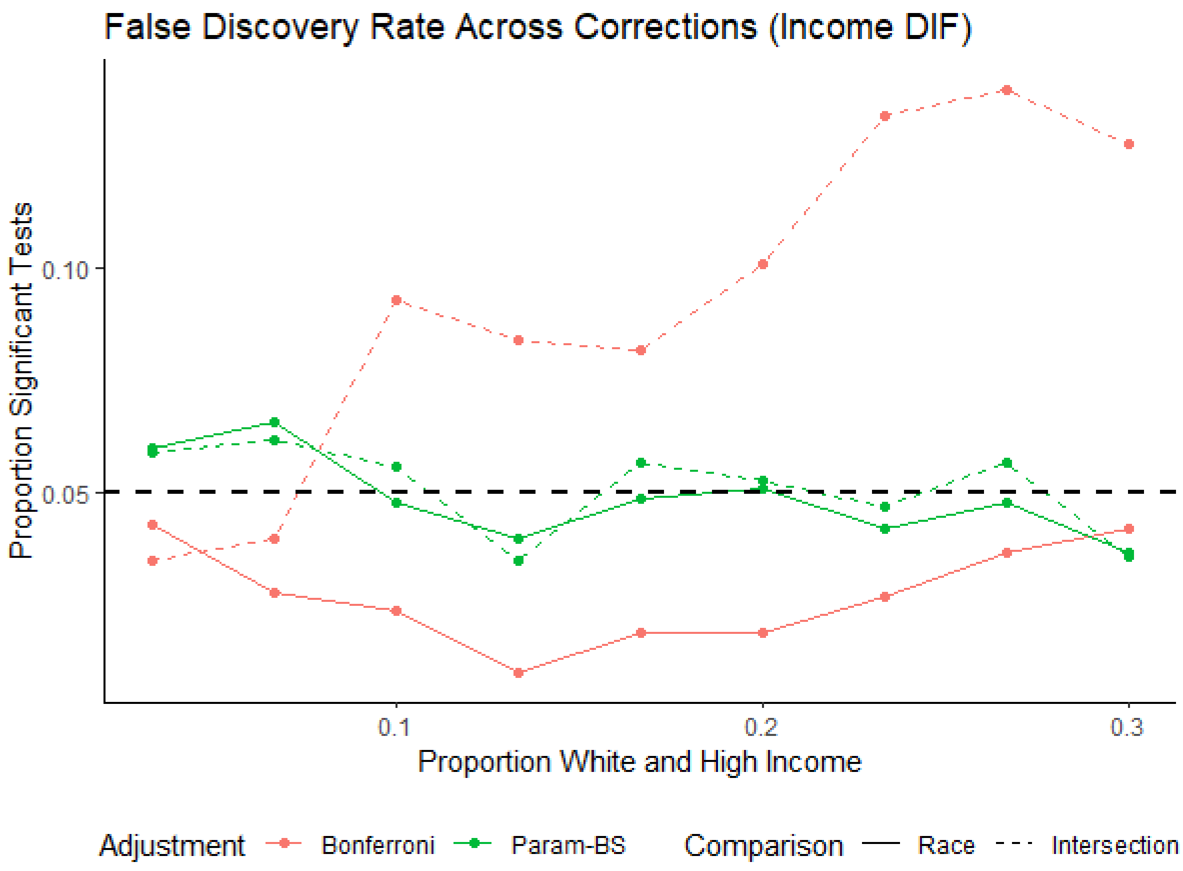

Figure 4. The proportion of incorrect DIF identification on non-DIF-containing comparisons across demographic distributions. Dashed line represents the critical value.

Note that the displayed results in Figure 4 include results from every simulated data set, including those where income was not initially detected as a source of DIF by the original MH test. Type II error, whereby a true effect goes undetected, can occur naturally; even though our simulations consistently impose a DIF effect of 0.3 in favor of high-income examinees, they still may not demonstrate substantial outperformance in some trials due to random chance. As such, the set of simulations where DIF is flagged as significant represents the set where DIF has manifested relatively strongly.

Figure 5 below recreates Figure 4 only for simulations where the DIF was significant in the MH test. On average, the false discovery rate for the parametric bootstrapping correction falls below the dashed five percent line, demonstrating a low false discovery rate for the parametric bootstrapping correction. Though the result is still a marked improvement over the Bonferroni correction, a researcher could bias their results by actively selecting for the cases where the income DIF is more pronounced before conducting parametric bootstrapping. This tends to overestimate the income DIF parameter used in parametric bootstrapping, and results in an improperly high expectation of the outperformance necessary on race or intersectional comparisons to qualify as significant given the estimated DIF on income. Though this issue is not unique to the present context, given that identifying significant results is the entire purpose of DIF detection, deliberate attention should be paid. We expound on this concern and others in the following section.

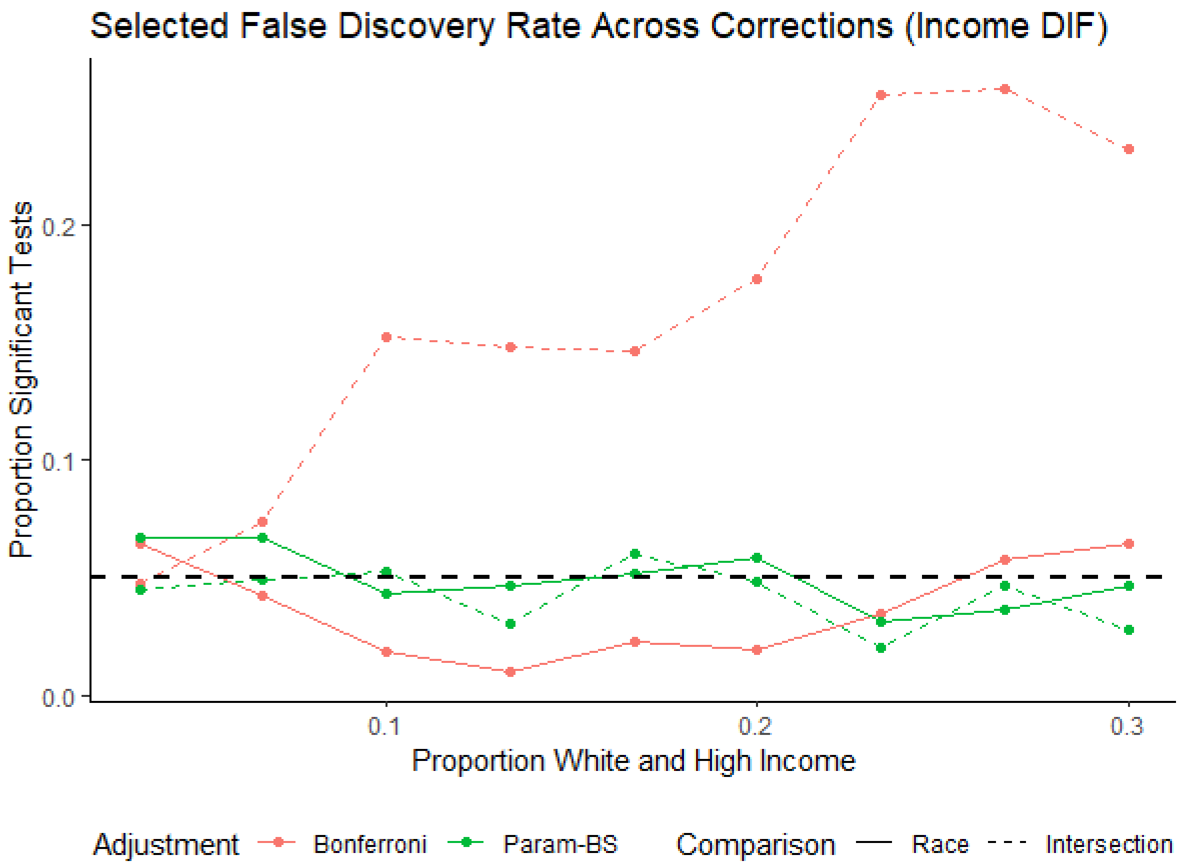

Figure 5. The proportion of spurious DIF identification based on biased parametrically-bootstrapped DIF parameters across demographic distributions. Dashed line represents the critical value.

5. Discussion

The findings of this study demonstrate that strategic implementation of parametric bootstrapping can properly adjust DIF detection processes by benchmarking expectations of observed statistics to empirically generated distributions. In this section we note some theoretical clarifications, implications, and limitations.

First, using parametric bootstrapping to correct for signal interference requires the research team to first declare a source for the observed DIF which, as we illustrated in this study, can be unclear especially in the case of high demographic overlap. At the same time, as shown in Figure 5 of the previous section, basing this choice on the significance of initial results can affect the validity of subsequent results. A common solution is an analysis plan (Olken, 2015): by a priori declaring the strategy to be implemented, the research team avoids biasing the final results. Another potential strategy is to acknowledge this threat and moderate conclusions in recognition of the potential bias. Future studies may discover a process for sequencing DIF tests without introducing bias.

A second, related, issue for signal interference occurs wherein effect A is accepted as true and effect B is suspected. The proposed method has a valid Type I error rate if effect B is zero, but under-identifies significant results for B otherwise. Consider where the demographics White and high-income have high overlap and both are true sources of DIF biased towards those identities. According to a typical analysis plan, if we found significant results for "White", we would use a logistic regression to estimate the DIF to be used for "White" in subsequent parametric bootstrapping. But the effect of "high-income" would be at least partially absorbed into the estimate for "White", and subsequent testing for a "high-income" effect would be based on just the remainder. This is similar to the issue of multicollinearity in regression models, where the estimates of two distinct effects become conflated. Traditional DIF testing is analogous to estimating each effect separately without regard to their overlap. Though one could include multiple demographic variables in the regression equation, this essentially elevates logistic regression to the DIF detection method, rather than, as in our examples, MH.

Third, the issues of 'correction overlap' and 'signal interference' are not unique to the context of intersectional DIF analysis. Even traditional DIF studies consider multiple demographic axes, and though these demographics are rarely independent, the p-value corrections employed typically ignore this nuance. Further, we suspect that in many previous studies where, say, both race and income were found to be significant, properly controlling for one would show the other loses significance. Whether there exists a substantive difference between race being a true source of DIF, and simply having high overlap with income which is a true source of DIF, is a discussion beyond the scope of this analysis, which only offers a method for mathematically separating sources. One notable exception is DIF detection methods that can test multiple sources simultaneously, though these present their own challenges, particularly with regard to sample size. Insofar as the MH test and others that test one dimension at a time remain standard practice, our parametric bootstrapping method will remain topical.

Fourth, a fully intersectional DIF perspective may seem to preclude the issues explored in this investigation. If, for example, to examine race and gender, one strictly separates examinees into “White-male”, “White-female”, “Black-male”, and “Black-female”, there would be no overlap in samples across DIF tests (source of overcorrection), and each effect would be considered independently (limiting the source of signal interference). But this ignores the core complexity of intersectionality. If one wished to then include income, the same issues arise: the race-by-gender-by-income results would shadow the original race-by-gender results, parallel to how in our simulation study the race-by-income results shadowed the income results. Instead of simulating the results of a triple interaction within a double interaction, we limited our study’s perspective to just two main effects and one subsequent interaction effect, but the theory is infinitely applicable. The alternative is to always begin DIF analyses interacted to the fullest extent possible given the available data, a standard and limitation which differs across study contexts.

The method proposed in this investigation naturally raises questions for future investigations. As noted, our simulation design engendered a disconnect between the DIF detection model (MH) and the parametric bootstrapping model (logistic regression) for the sake of simplicity, but there may be additional efficiency to be wrought in aligning the two stages. Further, parametric bootstrapping is not the only semi-parametric modeling tool. Early iterations of the present study produced preliminary evidence that a Westfall-Young procedure (Westfall & Young, 1993) had comparable performance, which we ultimately dropped to preserve the scope of the investigation. Methodological innovations in DIF detection as well as in statistical resampling techniques will inform next steps for the use of parametric bootstrapping in addressing statistical complications of intersectional DIF analyses.

6. Conclusion

Integrating intersectionality into DIF detection methods aligns psychometric tools with our increasingly complex notions of identity and systemic bias. Implementation seems to be straightforward as well: insofar as intersectionality subdivides demographic identities, examinee samples can be subdivided accordingly and tested for DIF along these more specific axes.

But this inherently raises methodological challenges, which we identify and demonstrate in this study: overcorrection of p-values with multiple testing adjustments and DIF signal interference between identity groups. The same sets of examinees will appear across multiple DIF tests according to the sample’s demographic distribution along multiple axes, which can obscure true DIF signals and distort error rates in ways that evade typical multiple hypothesis testing adjustments. These concerns demonstrate the challenge of balancing DIF detection power with rigorous control of error rates.

Our empirically validated solution to both complications involves parametric bootstrapping: using parameters routinely estimated during DIF analyses, we can simulate the distributions of relevant test statistics and subsequently adjust benchmarks to the desired null hypotheses. This strategy, by construction, exactly replicates the complex interdependence of any DIF tests, achieving the intended Type I error rate without sacrificing power under a needlessly conservative correction or necessitating derivation of a new closed form solution for every sample's exact dependency structure. Parametric bootstrapping accurately and efficiently identifies problematic items, allowing test developers to more efficiently respond to DIF.

Notably, the issues we identify are not exclusive to the given context. In traditional DIF detection the overlap of samples across tests is more subtle but similarly subject to the distortions demonstrated in the present study. Even from a fully intersectional DIF perspective, where DIF is only tested across intersected identities, any inconsistency in the level of interaction creates a parallel situation. The set of tests, the data generating parameters, and the specific invocation of parametric bootstrapping in the present analysis were chosen for illustrative clarity, but as an example framework this method has wide-reaching relevance and high potential for further innovations on power and accuracy.

DIF detection methods will continue to evolve alongside a deepening understanding of what DIF represents. We hope this work documenting the statistical properties of intersectional DIF opens the door for bolder exploration of this topic, which will ultimately improve fairness in testing.

7. References

Ackerman, T. A., & Ma, Y. (2024). Examining Differential Item Functioning from a Multidimensional IRT Perspective. *Psychometrika*, *89*(1), 4–41. https://doi.org/10.1007/s11336-024-09965-6

Albano, T., French, B. F., & Vo, T. T. (2024). Traditional vs Intersectional DIF Analysis: Considerations and a Comparison Using State Testing Data. *Applied Measurement in Education*, *37*(1), 57–70. https://doi.org/10.1080/08957347.2024.2311935

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Aryadoust, V., Min, S., & Chen, X. (2024). Investigating differential item functioning across interaction variables in listening comprehension assessment. *Studies in Educational Evaluation*, *80*, 101322. https://doi.org/10.1016/j.stueduc.2024.101322

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*(3), 507–526. https://doi.org/10.1037/met0000077

Bauer, D. J. (2023). Enhancing measurement validity in diverse populations: Modern approaches to evaluating differential item functioning. *British Journal of*

Mathematical and Statistical Psychology, *76*(3), 435–461. https://doi.org/10.1111/bmsp.12316

Bauer, D. J., Belzak, W. C. M., & Cole, V. T. (2020). Simplifying the Assessment of Measurement Invariance over Multiple Background Variables: Using Regularized Moderated Nonlinear Factor Analysis to Detect Differential Item Functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 43–55. https://doi.org/10.1080/10705511.2019.1642754

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni Del R Istituto Superiore Di Scienze Economiche e Commericiali Di Firenze*, *8*, 3–62.

Chalmers, R.P. (2012). "mirt: A Multidimensional Item Response Theory Package for the R Environment." *Journal of Statistical Software*, **48**(6), 1–29. doi:10.18637/jss.v048.i06

Chen, Y., Li, C., Ouyang, J., & Xu, G. (2023). DIF Statistical Inference Without Knowing Anchoring Items. *Psychometrika*, *88*(4), 1097–1122. https://doi.org/10.1007/s11336-023-09930-9

Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, *43*(6), 1241–1299. https://doi.org/10.2307/1229039

Debelak, R., & Debeer, D. (2024). mstDIF: A Collection of DIF Tests for Multistage Tests. R package version 0.1.8, https://CRAN.R-project.org/package=mstDIF

Dorans, N. J., & Holland, P. W. (1992). Dif Detection and Description: Mantel-Haenszel and Standardization1,2. *ETS Research Report Series*, *1992*(1), i–40. https://doi.org/10.1002/j.2333-8504.1992.tb01440.x

Else-Quest, N. M., & Hyde, J. S. (2016). Intersectionality in Quantitative Psychological Research: II. Methods and Techniques. *Psychology of Women Quarterly*, *40*(3), 319–336. https://doi.org/10.1177/0361684316647953

Holland, P. W., & Thayer, D. T. (1986). Differential Item Functioning and the Mantel-Haenszel Procedure. *ETS Research Report Series*, *1986*(2), i–24. https://doi.org/10.1002/j.2330-8516.1986.tb00186.x

Kim, J., & Oshima, T. C. (2013). Effect of Multiple Testing Adjustment in Differential Item Functioning Detection. *Educational and Psychological Measurement*, *73*(3), 458–470. https://doi.org/10.1177/0013164412467033

Mantel, N., & Haenszel, W. (1959). Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *JNCI: Journal of the National Cancer Institute*, *22*(4), 719–748. https://doi.org/10.1093/jnci/22.4.719

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, *29*(3), 61-80.

Penfield, R. D. (2001). Assessing Differential Item Functioning Among Multiple Groups: A Comparison of Three Mantel-Haenszel Procedures. *Applied Measurement in Education*, *14*(3), 235–259. https://doi.org/10.1207/S15324818AME1403_3

Russell, M. (2023). *Systemic Racism and Educational Measurement: Confronting Injustice in Testing, Assessment, and Beyond*. Routledge. https://doi.org/10.4324/9781003228141

Russell, M., & Kaplan, L. (2021). An Intersectional Approach to Differential Item Functioning: Reflecting Configurations of Inequality. *Practical Assessment, Research & Evaluation*, *26*(21).

Russell, M., Szendey, O., & Kaplan, L. (2021). An Intersectional Approach to DIF: Do Initial Findings Hold across Tests? *Educational Assessment*, *26*(4), 284–298. https://doi.org/10.1080/10627197.2021.1965473

Russell, M., Szendey, O., & Li, Z. (2022). An Intersectional Approach to DIF: Comparing Outcomes across Methods. *Educational Assessment*, *27*(2), 115–135. https://doi.org/10.1080/10627197.2022.2094757

Self, S. G., & Liang, K.-Y. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions. *Journal of the American Statistical Association*, *82*(398), 605–610. https://doi.org/10.1080/01621459.1987.10478472

Shaffer, J. P. (1995). Multiple Hypothesis Testing. *Annual Review of Psychology*, *46*(1), 561–584.

Sireci, S. G. (2020). Standardization and UNDERSTANDardization in Educational Assessment. *Educational Measurement: Issues and Practice*, *39*(3), 100–105. https://doi.org/10.1111/emip.12377

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *The Journal of Applied Psychology*, *91*(6), 1292–1306. https://doi.org/10.1037/0021-9010.91.6.1292

Suk, Y., & Han, K. C. T. (2023). *Evaluating Intersectional Fairness in Algorithmic Decision Making Using Intersectional Differential Algorithmic Functioning*. https://doi.org/10.31234/osf.io/e93js

Suk, Y., & Han, K. T. (2024). A Psychometric Framework for Evaluating Fairness in Algorithmic Decision Making: Differential Algorithmic Functioning. *Journal of*

Educational and Behavioral Statistics, *49*(2), 151–172. https://doi.org/10.3102/10769986231171711

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. https://doi.org/10.3102/10769986027001053

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley & Sons.