Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 26 Number 21, August 2021 ISSN 1531-7714

# An Intersectional Approach to Differential Item Functioning: Reflecting Configurations of Inequality

Michael Russell, *Boston College*
Larry Kaplan, *Boston College*

Differential Item Functioning (DIF) is commonly employed to examine measurement bias of test scores. Current approaches to DIF compare item functioning separately for select demographic identities such as gender, racial stratification, and economic status. Examining potential item bias fails to recognize and capture the intersecting configurations of inequality (McCall, 2001) specific to a person's identify which impact item bias. The study presented here explores an intersectional approach to the flagging of items for content review using the standardized-D DIF method. The intersectional approach aims to capture the confounding/compounding impacts of intersectional configurations of inequality.

## Introduction

The Joint Standards on Educational and Psychological Testing (AERA/APA/NCME, 2014) establish that fairness is a critical issue to examine when considering the use of a test, particularly when making high-stakes decisions at the individual level. Measurement bias is one factor that impacts fairness. Measurement bias occurs when a construct-irrelevant factor produces error in scores that is not random. Typically, construct-irrelevancy is associated with specific design aspects of an item and/or the conditions under which a test is administered. An example of the former is an item that employs an idiom that is unfamiliar to students from a specific geographic area. Administering a writing test on computers for students who are not accustomed to writing on computer is an example of the latter.

The Joint Standards recommend *differential item functioning* as one approach to examining potential bias in test scores. As the Joint Standards state, "differential item functioning (DIF) is said to occur when equally able test takers differ in their probabilities of answering a test item correctly as a function of group membership" (2014, p. 15). Holland (2008) emphasizes, however, that it is not group membership that causes an outcome, such as an item to behave differently. Rather, it is the lived experience, and often differences in how members of a group are treated by or interact with society, that produce differences in how characteristics of an item interplay with members of a group. In this way, differential item functioning implicitly aims to assure that the functioning of an item does not reflect bias, discrimination, or other forms of oppressive policies and actions that operate in society and adversely affect members of a focal group.

Several techniques for conducting a DIF analysis have been developed (Holland & Wainer, 1993; Osterlind & Everson, 2009). Although the methods used to calculate the statistic interpreted to indicate the potential presence of DIF differ across methods, there are seven features that are common across methods. The first feature is a focus on the functioning of each individual item rather than the test as whole. Second is the comparison of an item's functioning between two groups, one termed the reference group and the other termed the focal group. The question explored through this comparison is whether the item functions the same for the two groups conditioned on the ability of each group member. Third is the use of the total test score as a proxy for ability. Fourth is the formation of sub-groups based on test performance, or what researchers at the Educational Testing Service (ETS) term "slices" (Dorans & Holland, 1992). In effect, each slice represents a sub-group of test-takers of the same (or similar, depending on the range of the slice) ability. Fifth, the probability of responding correctly on the item of interest is compared between the reference and focal groups within each slice or the probability of responding correctly is estimated conditioned on assignment to a given slice. Sixth, a statistical method is applied to yield an overall estimate of the extent to which the probability of responding correctly to an item differed between the reference and focal group conditioned on ability. Finally, and perhaps most importantly, for all methods, the statistic produced is interpreted as an indicator of potential bias and serves as a trigger for additional analyses of the test content and/or test administration conditions for construct-irrelevant factors that may contribute to the item's differential functioning. In this way, a DIF statistic typically is not interpreted as an indicator of bias; at best, an indicator of DIF informs further investigation of potential bias.

For most testing programs, the sub-groups of interest typically include gender, race, economic status, second language status, and special education status. For gender, race, and economic status, the underlying question examined in a DIF analysis is whether the lived experience of people who identify with a given gender or a given racially stratified group, or live under different economic conditions, influences the probability of responding correctly to a given item. For second language status, the underlying question focuses on whether the language employed by a select item may differentially influence the probability of responding correctly for students whose first language is not the language in which the test is administered. For special education status, the underlying question examines whether features of the item differentially impact the item's ability to access the targeted construct due to different access needs between sub-group members.

## Confounding: A Challenge When Examining DIF for Multiple Demographic Characteristics

All DIF techniques contrast a reference group with a focal group. The reference group is typically identified as that group most advantaged in our society. When examining DIF by gender, Males are typically assigned the reference group, and females the focal group. When examining DIF by racial stratification, test-takers identified as White are typically the refence group, and members of each racially stratified group of interest form the focal group. For economic status, test-takers who are not identified as economically disadvantaged (and thus are advantaged) are typically assigned the reference group, and test-takers identified as economically disadvantaged are the focal group. For second language, test-takers whose first language is English or who are not receiving English language learning instruction are assigned to the reference group and test-takers whose first language is not English and/or who receive English language instruction form the focal group. And when examining DIF by special education status, test-takers who have not been identified with an individual education plan or in need of additional education services are assigned to the reference group and those who have an IEP or are identified as eligible for additional education services are the focal group.

Traditionally, separate DIF analyses are performed for each pair of sub-groups of interest. As an example, if both gender and economic status are of interest, one DIF analysis focuses on potential differences in item functioning by gender and a separate analysis focuses on potential differences by economic status. In such cases, a given item might be flagged for further investigation due to a potential difference identified between gender categories, economic status categories, or both.

In the traditional approach, a potential limitation results from the assignment of a given student to both a reference group and a focal group, depending on the demographic characteristic of focus. As an example, when DIF is examined separately for both gender and economic status, a portion of the male reference group contains members of an advantaged economic group and the remaining portion are members of the economically disadvantaged group, and the same confounding occurs for test-takers identified as female. Similarly, a portion of the economically advantaged reference group consists of males and the remaining portion is formed by females, as also occurs for economically disadvantaged test-takers. This confounding results in some test-takers being assigned to the reference group in both analyses (e.g., economically advantaged males), some test-takers assigned to the focal group in both analyses (e.g., economically disadvantaged females), and some test-takers assigned to the reference group in one analysis and the focal group in the other (e.g., economically advantaged females and economically disadvantaged males).

As an example, inequities produced by racism have placed a larger percentage of people who are Black in economic disadvantaged states (Oliver & Shapiro, 1989, 2001, 2006; Rothstein, 2017; Gillborn, 2010). As a result, there is potential for DIF associated with economic status to interact with DIF associated with Black racial stratification. Further, there is potential for DIF associated with both economic status and Black racial stratification to be relatively small, such that further review of an item is not warranted. Yet, the compound effect of oppression, or what McCall (2001, p. 6) terms “configurations of inequality,” experienced by test-takers who are both economically disadvantaged and Black may warrant further investigation.

## Intersectionality and DIF

To address the potential confounding of differential item functioning across categories of identity the study presented here explores the use of an intersectional approach to DIF analyses. Over the past decade, both Quantitative Critical Race Theory (QuantCRT) (Baker, 2019; Garcia et al., 2018; Gillborn et al., 2018) and Critical Quantitative Inquiry (CQI) (Denzin, 2017; Stage & Wells, 2014) have emphasized the importance of approaching analyses intended to explore differential effects or outcomes among sub-groups in a manner that reflects the intersections of each individual’s identities (Crenshaw 1991; Hancock, 2013; LaVeist, 1994; McCall 2001, 2005; Museus & Griffin 2011; Zuberi 2001).

Intersectionality recognizes that each person has multiple identities and that it is the nexus of these identities that influences their lived experiences (Crenshaw, 1991; McCall, 2001; Lopez et al., 2018). As an example, rather than possessing three distinct types of identities—gender, racially stratified group membership, and economic status—and understanding each as having separate and distinct influences on a person’s lived experiences, intersectionality recognizes that a person’s lived experiences are influenced by the intersection of these categories of identity. In an intersectional framework, a person is not understood as only male or female, Black or White, economically advantaged or disadvantaged. Instead, each person is recognized as a composite of these identities. A person is a female who is White and is economically disadvantaged or a male who is Black and is economically advantaged. And it is the intersection of these identities that impacts the cumulative effect of oppression or advantage associated with each identity.

When applied to DIF, an intersectional approach enables a single reference group to be defined by a select intersectional group. Each remaining intersectional group then forms a focal group that is compared to the same reference group. In doing so, the magnitude of a DIF statistic can be directly compared among focal groups because each statistic is in reference to the same reference group. In addition, interactions that may exist among traditionally defined demographic characteristics are accounted for directly within the focal groups. In turn, confounding DIF effects that may occur across demographic characteristics of interest are eliminated.

As one example, traditional analyses of the higher education pipeline typically compare entry and completion rates in at least three ways: people identified as white vs. people identified as of color (or sometimes specific sub-groups of people of color); people identified as male vs. people identified as female; and first-generation students vs. second generation and beyond students (Chapa & Schink, 2006; Horn, 1997; King, 2000; Mazon & Ross, 1990). These analyses consistently suggest that students identified as White experience higher levels of successful completion than students identified as of color, that female students typically have higher completion rates than males but that the magnitude of the difference is smaller than for the race/ethnicity comparison, and that first-generation students experience less success completing higher education than the reference group. In effect, this approach to examining the higher education pipeline separately for a given demographic characteristic parallels the approach typically taken for DIF analyses.

Intersectional analyses of the higher education pipeline take a different approach that begins by grouping students based on three or more demographic variables. As an example Lopez and her colleagues employed an intersectional approach in which each student's gender, economic status, and racially stratified identities were combined to represent their intersectional identity (Lopez et al., 2018). In their analysis students identified as female, from low-income households, and Black formed one group. Students identified as male, from high-income households and White formed a second group, and so on. Completion rates were examined separately for each group and the group with the highest completion rate served as the reference group in subsequent analyses. This procedure allowed all analyses to express differences among intersectional groups in reference to a single group. For higher education pipeline analyses, Lopez et al. (2018) defined the reference group as students identified as female, White, and from high-income households. As they summarize, this intersectional approach revealed "surprising race–gender–class gaps between both high- and low-income quartiles that would ordinarily remain unseen in conventional race-only, gender-only, and class-only reporting on graduation rates and developmental class placement" (2018, p. 181). What is attractive about this approach is that it reveals interesting, and previously undetected, differences that occur for specific intersectional subgroups that are masked by the multiple sub-group analyses.

The study presented below was conducted to explore the utility and potential challenges to employing an intersectional approach to examine differential functioning of items. It is important to emphasize that the study was limited to examining the impact an intersectional approach might have on identifying (aka, flagging) items in need of further review. The study did not proceed with a full review of any items. In a full DIF study, further review often fails to identify a potential cause for differential functioning and the item is deemed to be acceptable for operational use. Because we did not conduct a full review of flagged items, the findings presented below should not be interpreted as indicating the number of biased items. The counts and percentages presented only represent items flagged for further review. We focus our analysis only on the flagging of items for two reasons. First, flagging is the first, and arguably the most critical step, in identifying potentially biased items for the simple reason that if an item is not flagged, then no further consideration of potential item bias occurs. Second, content review requires access to the actual content of the items as well as participation by a panel of experts knowledgeable about the characteristics of an item that may disadvantage test-takers with specific life experiences. We did not have access to all of the items, because several remain secure, and, to date, panels with expertise in intersectional lived experiences have not been assembled and thus were not available to us. Finally, we opt not to name the state from which the test scores originate in order to protect the state from potential accusations of test bias based on only a partial analysis of item bias.

## Study Design and Analytic Methods

To examine the potential utility of an intersectional approach to forming reference and focal groups, this study conducted two sets of DIF analyses using results from students performance on a state's grade 5 operational English Language Arts (ELA) test. The data set was provided by the state and contained test and item scores for all students who performed the state ELA test. Demographic data, including gender and racially stratified identity was provided to the state by each school district and was originally collected from each student's parents/guardians at the time of enrollment in the district. Economic disadvantage was also provided to the state by each district and was defined based on participation in one of four programs designed to support students in households whose income is at or below 130% of the federal poverty guidelines.

The test contained 25 items, was designed to assess achievement of the state's ELA standards which were adapted from the Common Core State Standards, and had a score reliability (Cronbach's alpha coefficient) of

.90. For this test, DIF was examined using the standardized D-static method (Dorans & Kulich, 1986) for which the following formula is applied to calculate an indicator of potential DIF:

$$D_{std} = \frac{\sum_{s=1}^{S} K_s \left[ P_{fs} - P_{rs} \right]}{\sum_{s=1}^{S} K_s}$$

Where:

$P_{fs}$ is the percent correct for the focal group for students in ability band s

$P_{rs}$ is the percent correct for the reference group for students in ability band s

$K_s$ is the weight for ability band s

There are four different approaches to calculating $K_s$:

1. The number of people at s in the total group, $N_{ts}$
2. The number of people at s in the reference group, $N_{rs}$
3. The number of people at s in the focal group, $N_{fs}$
4. The relative number of people in some standard reference group, for example a 3-year rolling norms group for the SAT

For our analysis, $K_s = N_{fs}$ was used. As Dorans & Kulich (1986) note, this approach gives the greatest weight to differences in $P_{fs}$ and $P_{rs}$ at those ability levels most attained by the focal group and is the approach typically practiced.

The following criteria were applied to determine whether potential DIF occurred and, if so, whether DIF was suspicious or likely (Dorans et al., 1992):

| | |
|---|---|
| $.00 \leq \text{abs}(D_{std}) < .05$ | No DIF |
| $.05 \leq \text{abs}(D_{std}) < .10$ | Suspicious |
| $.10 \leq \text{abs}(D_{std})$ | Likely |

It should be noted that most large-scale testing program only review items for which the standardized D static exceeds .10. As noted above, several methods for examining DIF have been developed. We opted to use the standardized-D statistic for two reasons. First, the state from which the data come employs the standardized-D statistic which allowed us to confirm that the findings from our analyses using the traditional approach to forming reference and focal groups was consistent with the state's findings. Second, both standardized-D and logistic regression are employed by several state testing programs to examine DIF. Logistic regression, however, has been shown to be unstable in detecting DIF for both large sample sizes and when there are large differences between the size of the sample for the reference and focal groups (Cuevas & Cervantes, 2012). Because our analyses aimed to compare findings across methods, we believed it was important to use the same set of students for all analyses. This resulted in large samples for some groups, in some cases exceeding 30,000, and large differences in sample sizes between groups, in the extreme a difference exceeding 20-fold.

To calculate the standardized-D statistic, test-takers were first categorized into ability bands based on their total test score. The grade 5 ELA test employed a scaled score that had a 120-point range. A scale score range of 10 was selected to define each ability-level slice such that 12 slices mere formed. We opted to use a 10-point range in order to ensure a minimum of 10 students in each band for all intersectional groups.

Two approaches were applied to form the reference and focal groups. For both methods, three demographic characteristics were of interest, namely gender (male/female), racial stratification (White/Black/Hispanic/Asian), and economic status (Economic Advantaged/Economic Disadvantaged). Only students for whom demographic data was reported for all three of these characteristics were included in the analyses. In addition, students whose racially stratified identity was something other than White, Black, Hispanic, or Asian were excluded from analysis because there were not enough students with these identities to form intersectional groups of sufficient size to conduct DIF analyses. In total, these criteria excluded 2.4% of the full population of test-takers from the analyses.

The first approach, which we term traditional, examined DIF separately for each demographic characteristic. For gender, Male was defined as the reference group and Female was the focal group. For racial stratification, White was defined as the reference group and Black, Hispanic, and Asian each formed a separate focal group. And for economic status, Economic Advantaged was defined as the reference group and Economic Disadvantaged formed the focal group. The traditional approach resulted in five sets of DIF statistics for each item, each comparing a given focal group to its corresponding reference group.

The second approach, which we term intersectional, combined the three demographic characteristics to form 16 intersectional groups listed in Table 1. Male-White-Advantaged was defined as the reference group. Because Male-White-Advantaged are viewed as the most advantaged group in U.S. society (i.e., the nation in which the state test that is the focus of analyses was administered), defining this intersectional group as the reference group is consistent with the logic employed in traditional DIF analyses. Each of the 15 remaining intersectional groups formed a focal group.

The total sample size for this study was approximately 67,000 test-takers. The sizes for intersectional groups varied considerably. The smallest two groups (F-A-D and M-A-D) contained approximately 600 students. The largest groups (M-W-A and F-W-A) contained approximately 16,000 students. Most groups, however, contained between 1,000 and 5,000 students. There was also considerable variation in the mean scale score among the intersectional groups. Test-takers identified as Female-Asian-Advantaged received the highest mean score (518) and students identified as Male-Hispanic-Disadvantaged received the lowest mean score (486). Test-takers identified as Male-White-Advantaged had the fifth highest mean score (504).

## Findings

To explore the use of intersectional groupings to examine differential functioning of the 25 items comprising the grade 5 ELA test, two sets of analyses were conducted. In this section, findings are presented separately for the two approaches. Findings are then compared across the two approaches.

### Traditional Groupings

DIF analyses typically compare item functioning by gender, race/ethnicity, and economic status, among other demographic characteristics. Table 2 presents the standardized D statistic for each of the 25 items on the Grade 5 ELA test for each focal group examined. Cells shaded light green indicate a positive standardized D statistic that meets the criteria for suspicious DIF. Dark green shading indicates a positive standardized D statistic that meets the criteria for likely DIF. Light red shading indicates a negative standardized D statistic that meets the criteria for suspicious DIF. And dark red shading indicates a negative standardized D statistic that meets the criteria for likely DIF. The final two columns indicate the number of focal groups for which an item was flagged as suspicious or likely DIF. The final two rows indicate the number of items for a given focal group that were flagged as suspicious or likely DIF.

Table 1. Intersectional Groups

| Group | Code | Group | Code |
|---|---|---|---|
| Male-White-Advantaged | MWA | Male-White-Disadvantaged | MWD |
| Male-Black-Advantaged | MBA | Male-Black- Disadvantaged | MBD |
| Male-Hispanic-Advantaged | MHA | Male-Hispanic- Disadvantaged | MHD |
| Male-Asian-Advantaged | MAA | Male-Asian- Disadvantaged | MAD |
| Female-White-Advantaged | FWA | Female-White- Disadvantaged | FWD |
| Female -Black-Advantaged | FBA | Female -Black- Disadvantaged | FBD |
| Female -Hispanic-Advantaged | FHA | Female -Hispanic- Disadvantaged | FHD |
| Female -Asian-Advantaged | FAA | Female -Asian- Disadvantaged | FAD |

When examined by focal group, Table 2 indicates that two items were flagged as suspicious for females, one item was flagged as suspicious for test-takers identified as economically disadvantaged, one item was flagged as suspicious for test-takers identified as Black, two items were flagged as suspicious for test-takers identified as Hispanic, and two items were flagged as suspicious for test-takers identified as Asian. In all cases, the standardized D statistic is negative which indicates the item was more difficult for the focal group than for the reference group. Also note that, of the items flagged as suspicious, four were flagged for only one focal group and one item was flagged for four of the five focal groups.

Table 2. Grade 5 ELA Standardized D Statistics for Traditional Groupings

| Item | Female | Economic Disadvantaged | Black | Hispanic | Asian | Suspicious | Likely |
|---|---|---|---|---|---|---|---|
| 1 | -.05 | -.02 | -.03 | -.03 | -.04 | 1 | 0 |
| 2 | -.03 | -.02 | -.04 | -.04 | -.01 | 0 | 0 |
| 3 | -.02 | -.02 | -.05 | -.04 | -.02 | 0 | 0 |
| 4 | -.01 | -.02 | -.05 | -.04 | -.08 | 1 | 0 |
| 5 | -.02 | .00 | -.02 | -.01 | -.02 | 0 | 0 |
| 6 | -.01 | -.06 | -.07 | -.08 | -.06 | 4 | 0 |
| 7 | -.02 | .00 | -.02 | -.02 | .04 | 0 | 0 |
| 8 | .03 | .01 | .01 | .00 | -.01 | 0 | 0 |
| 9 | .01 | .02 | .02 | .03 | .00 | 0 | 0 |
| 10 | .00 | -.01 | .00 | -.02 | -.01 | 0 | 0 |
| 11 | .01 | -.03 | -.03 | -.03 | -.02 | 0 | 0 |
| 12 | .00 | -.01 | -.04 | -.03 | -.04 | 0 | 0 |
| 13 | .01 | -.04 | .00 | -.05 | -.01 | 0 | 0 |
| 14 | -.06 | -.02 | -.03 | -.03 | -.04 | 1 | 0 |
| 15 | .00 | .01 | .02 | .01 | .00 | 0 | 0 |
| 16 | .01 | .01 | .00 | .01 | .01 | 0 | 0 |
| 17 | .03 | .01 | .03 | .03 | .03 | 0 | 0 |
| 18 | -.01 | -.01 | .01 | -.01 | .00 | 0 | 0 |
| 19 | -.03 | -.01 | -.03 | -.02 | -.01 | 0 | 0 |
| 20 | -.01 | .00 | -.01 | -.01 | .01 | 0 | 0 |
| 21 | -.03 | -.02 | .01 | -.01 | .02 | 0 | 0 |
| 22 | -.04 | -.01 | .00 | .00 | -.01 | 0 | 0 |
| 23 | .00 | -.04 | -.04 | -.06 | -.01 | 1 | 0 |
| 24 | -.01 | .00 | -.03 | -.03 | -.02 | 0 | 0 |
| 25 | .01 | .01 | .02 | .03 | .04 | 0 | 0 |
| Suspicious | 2 | 1 | 1 | 2 | 2 | 8 | |
| Likely | 0 | 0 | 0 | 0 | 0 | | 0 |

Note: Due to rounding, some cells report a Standardized D statistic of .05 and are not highlighted.

It is important to emphasize that the criterion for reviewing items typically requires standardized D to exceed .10. Based on the information presented in Table 2, no items meet this criterion and thus no items would require further review.

## Intersectional Groups

The intersectional method categorized test-takers based on the intersection of three demographic characteristics: gender, racial stratification, and economic status. There were two gender groups (Male and Female), four racially stratified groups (White, Black, Hispanic, and Asian), and two economic status groups (economically advantaged and economically disadvantaged). Categorizing test-takers based on the intersection for these three demographic characteristics results in 16 intersectional groups.

Table 3 presents findings for the standardized D DIF analyses with test-takers identified as Male-White-Advantaged defined as the reference group. Across the 25 items comprising the ELA test and the 15 focal groups, a total 375 comparisons were made. In total, 63 comparisons resulted in a standardized D statistic that met the criteria for suspicious DIF and 10 comparisons met the criteria for likely DIF. The majority of comparisons that met either condition indicated DIF that favored the reference group (highlighted red), which indicates the item was harder for the focal group than the reference group.

Focusing on the items, 8 items were not flagged for any of the focal groups. One item (#6) was flagged for 10 of the 15 focal groups, one item (#1) was flagged for 9 focal groups, one item (#14) was flagged for 8 focal groups, two items were flagged for 6 focal groups (#3 & 4) and the remaining items were flagged for five or fewer focal groups.

Focusing on the focal groups, note that every group, except Male-White-Disadvantaged, was flagged for at least one item. The Female-Black-Advantaged intersectional group was flagged for 10 of the 25 items. The Female-Black-Disadvantaged, Female-Hispanic-Disadvantaged, and Female-Asian-Disadvantaged groups were each flagged for 9 items. Finally, the Male-Asian-Disadvantaged and Female-Hispanic-Advantaged groups were each flagged for seven items. All of the remaining intersectional groups were flagged for five or fewer items. It is noteworthy that of the nine items for which the Female-Asian-Disadvantaged group was flagged, four met the criteria for likely DIF.

It is also interesting to note that intersectional groups that contained females generally had more flags than the corresponding group that contained males. As an example, for students identified as Black and economically disadvantaged, females were flagged for 9 items while males were flagged for only one item. A notable difference, however, occurred for students identified as Asian and disadvantaged; both females and males had seven or more items flagged.

## Comparing Findings from Traditional and Intersectional Approaches

Comparing the traditional and intersectional approaches to framing DIF analyses reveals three noteworthy observations. First, forming groups based on the intersection of three demographic characteristics greatly increases the number of groups examined and thus the number of comparisons made. Whereas the traditional method focused on 5 focal groups (female, Black, Hispanic, Asian, and economically disadvantaged) which resulted in 125 item-level comparisons, the intersectional method focused on 15 focal groups which resulted in 375 item-level comparisons. Given the increased number of comparisons in the intersectional group method, it is likely that a larger number of comparisons will be flagged due to chance alone. However, across all methods, one would expect the percentage of comparisons flagged by chance to be similar. Table 4 shows the percentage of comparisons flagged for each method and indicates considerable differences among the approaches. Whereas the traditional method yielded only 6.4 percent of standardized D statistics as suspicious and none likely, the intersectional method yielded a notably higher percentage of comparisons resulting in flags.

A second noteworthy observation pertains to the number of items that were flagged for students with specific identities. The traditional method flagged a maximum of 2 items for any one focal group, all at the suspicious level. In contrast, the intersectional method flagged seven or more items for six of the intersectional groups. As noted above, the intersectional approach makes clear that the differential functioning of items compounds when one

Table 3. Male-White-Advantaged Intersectional Reference Group Standardized D Statistics

| Item | M-B-A | M-H-A | M-A-A | M-W-D | M-B-D | M-H-D | M-A-D | F-W-A | F-B-A | F-H-A | F-A-A | F-W-D | F-B-D | F-H-D | F-A-D | Suspi-cious | Likely |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -.04 | -.03 | -.03 | -.01 | -.03 | -.04 | -.08 | -.07 | -.07 | -.07 | -.11 | -.06 | -.06 | -.08 | -.13 | 7 | 2 |
| 2 | -.04 | -.03 | -.02 | -.01 | -.04 | -.05 | 0 | -.02 | -.07 | -.06 | -.03 | -.04 | -.07 | -.09 | -.04 | 4 | 0 |
| 3 | -.06 | -.04 | -.01 | -.01 | -.05 | -.05 | -.06 | -.01 | -.06 | -.05 | -.03 | -.03 | -.07 | -.07 | -.07 | 6 | 0 |
| 4 | -.07 | -.04 | -.08 | -.01 | -.04 | -.04 | -.12 | -.01 | -.06 | -.03 | -.08 | -.01 | -.04 | -.04 | -.12 | 4 | 2 |
| 5 | -.03 | 0 | -.03 | 0 | -.02 | 0 | -.02 | -.02 | -.04 | -.04 | -.04 | -.02 | -.03 | -.03 | -.03 | 0 | 0 |
| 6 | -.04 | -.06 | -.06 | -.02 | -.08 | -.11 | -.12 | -.01 | -.05 | -.05 | -.04 | -.03 | -.1 | -.1 | -.12 | 6 | 4 |
| 7 | 0 | -.03 | .06 | .02 | 0 | 0 | -.01 | -.02 | -.05 | -.05 | .02 | -.02 | -.04 | -.03 | .02 | 2 | 0 |
| 8 | .02 | .01 | 0 | .02 | .03 | .02 | .02 | .04 | .05 | .03 | .02 | .06 | .05 | .03 | .04 | 2 | 0 |
| 9 | .02 | .02 | 0 | .01 | .01 | .03 | .02 | .01 | .03 | .05 | .01 | .03 | .03 | .05 | .03 | 0 | 0 |
| 10 | .01 | -.02 | -.02 | -.01 | 0 | -.01 | 0 | -.01 | 0 | -.01 | -.02 | -.01 | 0 | -.02 | -.02 | 0 | 0 |
| 11 | -.04 | -.03 | -.02 | -.02 | -.04 | -.05 | -.08 | 0 | .01 | -.03 | 0 | -.02 | -.05 | -.03 | -.03 | 2 | 0 |
| 12 | -.05 | -.01 | -.05 | .01 | -.04 | -.02 | -.04 | 0 | -.03 | -.03 | -.04 | -.01 | -.05 | -.03 | -.06 | 2 | 0 |
| 13 | 0 | -.04 | -.01 | -.03 | -.02 | -.06 | -.05 | .01 | 0 | -.01 | 0 | -.02 | 0 | -.06 | -.05 | 2 | 0 |
| 14 | 0 | -.01 | -.01 | 0 | -.03 | -.02 | -.04 | -.05 | -.11 | -.09 | -.08 | -.05 | -.08 | -.09 | -.17 | 6 | 2 |
| 15 | .02 | 0 | -.02 | 0 | .02 | .03 | .03 | -.01 | .01 | 0 | -.01 | 0 | .02 | .01 | 0 | 0 | 0 |
| 16 | 0 | .01 | .01 | .01 | .01 | .03 | .04 | .01 | -.01 | .01 | .02 | .01 | .04 | .03 | .02 | 0 | 0 |
| 17 | .02 | .02 | .03 | -.01 | .01 | .02 | .03 | .03 | .04 | .06 | .06 | .02 | .05 | .05 | .08 | 5 | 0 |
| 18 | .01 | -.01 | .01 | -.01 | .02 | -.01 | .02 | 0 | -.01 | -.01 | -.01 | -.02 | .01 | -.02 | -.01 | 0 | 0 |
| 19 | 0 | 0 | 0 | 0 | -.02 | -.02 | -.02 | -.02 | -.07 | -.05 | -.02 | -.03 | -.08 | -.07 | -.04 | 4 | 0 |
| 20 | .01 | .01 | .02 | .02 | 0 | -.01 | 0 | 0 | -.02 | 0 | .02 | 0 | -.02 | -.01 | -.01 | 0 | 0 |
| 21 | .03 | .02 | .0 | 0 | 0 | -.01 | -.01 | -.02 | 0 | -.03 | 0 | -.04 | -.03 | -.05 | -.03 | 0 | 0 |
| 22 | 0 | 0 | .01 | 0 | 0 | 0 | 0 | -.03 | -.04 | -.04 | -.06 | -.05 | -.04 | -.04 | -.07 | 3 | 0 |
| 23 | -.03 | -.05 | 0 | -.03 | -.05 | -.06 | -.05 | .01 | -.02 | -.04 | 0 | -.02 | -.06 | -.07 | -.05 | 4 | 0 |
| 24 | -.04 | -.05 | -.0 | .01 | -.02 | -.01 | 0 | -.02 | -.07 | -.06 | -.01 | -.01 | -.03 | -.03 | -.05 | 2 | 0 |
| 25 | .01 | .02 | .04 | 0 | .02 | .03 | .05 | .01 | .03 | .03 | .05 | .01 | .03 | .04 | .07 | 2 | 0 |
| | | | | | | | | | | | | | | | | | |
| Suspicious | 3 | 1 | 3 | 0 | 1 | 2 | 5 | 2 | 9 | 7 | 4 | 4 | 9 | 8 | 5 | 63 | |
| Likely | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 4 | | 10 |

Note: Due to rounding, some cells report a Standardized D statistic of .05 and are not highlighted.

considers both one's gender and racially stratified identity. As shown in Table 3, for both economically advantaged and disadvantaged students, more than 25% of items are flagged for females identified as Black or Hispanic. And, although students identified as Asian are often stereotyped as academically successful (which is not to be confused with advantage), the intersectional approach indicates considerable variation in differential functioning of items across this sub-group. Most striking is the large number of items flagged for students identified as economically disadvantaged, female, and Asian; a group for which nine out of 25 items (36%) were flagged, with four of those items (16%) meeting the criteria for likely DIF.

A final observation focuses on the number of groups for which items were flagged. In the traditional analysis, a total of five items were flagged at the suspicious level and none at the likely level. Of these five items, four were flagged for only focal group, and one item (#6) was flagged for four of the five focal groups (all groups except female). In the intersectional method, 17 items were flagged for at least one group at the suspicious level and four items were flagged at the likely level for at least one group. Three of the items flagged as likely were flagged for two groups, while one item (#6) was flagged for four groups.

Table 4. Percentage of Comparisons Flagged by Method

| | **Standardized D** | |
|---|---|---|
| | **Suspicious** | **Likely** |
| Traditional | 6.4% | 0.0% |
| Intersectional | 16.7 | 2.7 |

Perhaps not surprising, the items flagged for several groups (5 or more) in the intersectional method were also flagged for at least one group in the traditional method. Focusing on the two items flagged in the traditional method for gender (#1 and 14), we see that the intersectional method flagged all focal groups that comprised females. For item 1, it is also interesting to note that the traditional method did not flag this item for test-takers identified as Asian, but the intersectional method flagged the item for three of the four intersectional groups that contained test-takers identified as Asian.

Examining item 4, the traditional method flagged the item at the suspicious level for students identified as Asian. The intersectional method also flagged each focal group comprising students identified as Asian, regardless of their gender or economic status. Of these groups, both males and females identified as economically disadvantaged were flagged at the likely level. It is interesting to note that while the traditional method only flagged item 4 for students identified as Asian, the intersectional method also flagged two groups composed of students identified as Black (male-advantaged and female-disadvantaged).

The traditional method flagged item 23 for students identified as Hispanic. The intersectional method also flagged this item for students identified as Hispanic, but only for those who are economically disadvantaged. In addition, the intersectional method flagged the item for two additional sub-groups of economically disadvantaged students, males identified as Asian and females identified as Black.

Item 6 was flagged for every focal group except females, suggesting the item was suspicious for all racially stratified groups and students who are economically disadvantaged. The intersectional method failed to flag any intersectional groups composed of students identified as White, regardless of their economic status. It is interesting to note that two groups containing students identified as people of color and economically advantaged were also not flagged (Male-Black-Advantaged and Female-Asian-Advantaged). All other intersectional groups were flagged. Of those groups flagged, both students identified as Hispanic or Asian who are economically disadvantaged, regardless of gender, were flagged at the likely level.

Two additional items warrant comment. Item 3 was not flagged for any focal groups by the traditional method, but was flagged as suspicious for six groups in the intersectional method. Of these flagged groups, all are composed of students identified as people of color, some of whom are economically advantaged and some that are not. Item 17 is also of note. Item 17 was flagged as suspicious for five focal groups in the intersectional method, all of which contained females identified as people of color, but whose economic advantaged varied. What is most interesting about this item is the direction of the potential DIF. Whereas the vast majority of items were flagged for negative DIF, item 17 was flagged for positive DIF for all five focal groups.

Taken together, the analyses presented above indicate that the method used to examine DIF matters. When the traditional approach was employed, the standardized D criterion for likely DIF resulted in no items being flagged for any reference group. When the intersectional method was used, four items were identified. In addition, comparing the few items flagged as suspicious by the traditional method, we see that for many of these items not all members of the traditionally-defined focal group were flagged by the intersectional method and that, in several cases, groups including students with other identified characteristics were also flagged. Collectively, these analyses suggest that the method employed to define group membership impact findings from a DIF analysis.

## Discussion

The many forms of oppression that operate within the United States to produce advantage for some people and disadvantage for others have and continue to contribute to differences in each person's lived experiences. It is the interactions between these differences in lived experiences and the content employed by each test item and/or the administrative conditions under which a test is administrated that holds potential to produce bias in the measurement of a cognitive construct. DIF is the most common approach employed to examine potential bias at the item level. Since its introduction more than forty years ago (Scheuneman, 1979; Lord, 1980), DIF analyses have focused on potential bias related to broad categories of oppression, including gender, racial stratification, economic class, and ableness.

More recently, efforts to examine the effects of oppression on various outcomes have recognized that the life experienced by an individual is a composite of their many identities (Crenshaw, 1991; McCall, 2005). To more fully represent a person's identity and capture the multiple, and often compound, impacts of oppression, an intersectional approach is necessary. The study presented here applied an intersectional approach to DIF analyses and compared the flagging of items for potential bias with the approach traditionally employed in DIF analyses to define group membership.

For this study, two methods were applied to form groups based on demographic characteristics. The traditional approach focused on three distinct demographic characteristics, namely gender, racial stratification, and economic status. Analyses focused on each demographically defined group separately. And for each demographic group, the dominant sub-group was defined as the reference group and the remaining group(s) were defined as the focal groups. The intersectional approach defined group membership based on the intersection of the same three demographic characteristics such that students were assigned to a sub-group based on the intersection of their identified gender, racial stratification, and economic status. For the intersectional approach, the dominant group in our society, namely males identified as White who are economically advantaged, was defined as the reference group and the remaining 15 intersectional groups each served as a focal group. For all DIF analyses, the standardized D method was employed and two criteria were applied to flag items; standardized D between |.05| and |.10| were flagged as suspicious and standardized D greater than |.10| were flagged as likely.

The findings indicate that the method employed to define group membership did affect the number and percentage of items flagged as both suspicious and likely. Whereas the traditional approach flagged five (20%) items as suspicious for one or more focal groups and no items as likely, the intersectional approach flagged 17 (68%) of the items as suspicious and four (16%) of the items as likely. A similar pattern also occurred when focusing on the number of items flagged for a given focal group. In the traditional approach, the maximum number of items flagged as suspicious for a given focal group was two. In contrast, in the intersectional approach, every group except students who were identified as male, White, and economically disadvantaged, had at least one item flagged, and six groups had seven or more items flagged. It should be noted that students identified as female, Asian and economically disadvantaged had the largest number of items flagged as likely (4).

### Practical Issues for Consideration

The study presented here provides preliminary evidence that an intersectional approach to defining reference and focal groups increases concerns test developers will likely have regarding potential bias in test items for test-takers with specific intersectional identities. The implementation of an intersectional approach, however, presents at least three practical challenges specific to sample sizes, multiple comparisons, and review of flagged items. Each of these topics is discussed separately below.

*Sample Size.* The study presented here used data from the full population of students who performed a state's grade 5 ELA test. For this state, approximately 98% of eligible students perform the state test. In our study, we included 97.6% of the students who took the test and had all three demographic characteristics of interest reported in the state data file, which resulted in a sample of over 60,000 test-takers. This relatively large sample of test-takers allowed us to form 16 intersectional groups the smallest of which contained nearly 600 students and the largest of which contained more than 16,000 students. Typically, DIF analyses are performed with field test data and contain much smaller sample sizes. Clearly, smaller sample sizes may create challenges for forming some intersectional groups that represent a smaller percentage of the total population of test takers.

Research has shown that sample size and differences between the sample size of the reference and focal group can impact DIF analyses. As an example, Cuevas and Cervantes (2012) found that, when employing logistic regression, large samples inflated the detection of potential DIF when statistical significance was employed to inform flagging of items for potential DIF. Large differences in sample size between the reference and focal group, however, resulted in under-flagging items when effect sizes were used to establish flagging criteria.

Recommendations on minimum sample sizes for DIF analyses vary. As an example, the Educational Testing Service states that "at least 200 members in the smaller group and at least 500 in total are needed for DIF analyses performed at the test assembly phase. For DIF analyses performed at the preliminary item analysis phase (after a test has been administered but before scores are reported), the minimum sample size requirements are 300 members in the smaller group and 700 in total" (Zwick, 2012, p. 11). Cognia (formerly Measured Progress), however, conducts DIF analyses for all subgroups with at least 75 students (Massachusetts Department of Education, 2018). And through a series of simulation studies, Belzak (2020) found that uniform DIF was detected with reasonable accuracy with samples as small as 50 per group.

The feasibility of conducting DIF with samples much smaller than that employed for this study suggests that, with careful sampling, an intersectional approach to reference and focal group formation is possible during field testing, particularly when field test items are embedded in operational test forms. This feasibility is particularly applicable to digitally-delivered tests for which test-taker demographic information is available prior to test administration. Whereas field testing often relies on random distribution of test forms, a program could capitalize on test-taker demographic information to stratify the random assignment of test forms within intersectional groups. This would allow a program to both define the number of people within each intersectional group that is administered a given form and assure adequate sample sizes for each form. Further, if the lower minimum thresholds employed by Cognia are adopted, the sample employed for this study would allow at least six field test forms to be administered while maintaining minimum samples of nearly 100 per item for even the smallest intersectional groups. Of course, the number of field test items that could be embedded in an operational test administration is impacted by the population of test-takers served by the testing program and the proportion of the total sample represented by a given intersectional group, both of which vary across states.

*Multiple Comparisons.* As noted above, the intersectional method examined here greatly increased the number of sub-group comparisons conducted. Whereas 125 comparisons were made when the traditional method was applied, 375 comparisons were made during the intersectional approach. The increased number of comparisons is expected to increase the number of flagged items simply by chance alone. When multiple statistical tests are conducted, researchers often adjust the alpha level and/or p-value used to determine statistical significance. It is interesting to note that a review of more than a dozen state testing program technical reports indicate that the current practice does not make adjustments for DIF analyses despite the multiple comparisons that occur when employing the traditional approach to defining reference and focal groups. Even when limited to racial stratification, analyses reported in technical reports defined students identified as White as the reference group and compared them separately to students identified as Black, Latinx/Hispanic, Asian, and, when sample sizes allow, American Indian/Alaska Native, Pacific Islander, as well as students identified as two or more races. Similarly, as testing programs have transitioned to digitally-delivered tests, DIF has been used to examine differential item functioning between various technological factors including screen size, screen resolution, browser type, and availability and/or use of specific accessibility features. For each of these comparisons, a given medium of test administration (e.g., paper-based or desktop computer) defines the reference group and each technological factor defines a reference group. As just one example, Oklahoma's (2019) technical report presents findings from ten technological factors and ten demographic characteristics. For the demographic characteristics, five sets of comparisons are made in which test-takers identified as White serve as the reference group. For the technological factors, three sets of comparisons are made in which test-takers using the Chrome operating system serve as the reference group. Despite these two sets of repeated comparisons, no adjustments are made to protect against false discovery.

Although current practice does not adjust for multiple comparisons, we acknowledge this is an issue that requires further consideration. For DIF analyses that employ logistic regression (Swaminathan & Rogers, 1990), the Benjamini-Hochberg (1995) procedure might be applied to adjust p-values to control for the false discovery rate during simultaneous or repeated inferences. It is interesting note, however, that at least one simulation study that examined the impact of sample size on DIF detection using logistic regression and which adjusted p-values to control for multiple comparisons found that doing so negatively impacted DIF detection as sample sizes decreased (Belzak, 2020). Other techniques for examining DIF, such as standardized D-statistic (Dorans & Kulich, 1986), root-mean-weighted squared difference (Dorans & Kulick, 1986), and Mantel-Haenszel (Fidalgo et al., 2004; Zwick, 2012) do not rely on tests of statistical significance as a criterion for identifying items for potential bias. As a result, adjustments to p-values to control for false discovery are not applicable. Again, although not current practice, one might nonetheless consider adjusting criteria for flagging to account for multiple comparisons. Simulation studies that manipulate the number of comparisons made are one approach to informing the development of adjustment procedures. In addition, procedures that simultaneously estimate differential functioning across multiple focal groups might be explored. As an example, Penfield (2001) explored the use of the Generalized Mantel-Haenszel statistic (Somes, 1986) to examine DIF simultaneously across four focal groups. Similarly, Shealy and Stout (1993) developed and applied SIBTEST to estimate DIF simultaneously across multiple items.

*Review of Flagged Items.* Although a review of the content of each flagged item was not conducted as part of this study, a potential challenge produced by the intersectional method is the type of expertise required to review flagged items. The traditional approach typically seeks experts familiar with content that may produce bias due to gender, racial stratification, or economic class. In most cases, different sets of people focus on each form of potential bias. For the intersectional approach, potential bias is flagged and believed to operate as a result of the intersection of one's gender, racially stratified identity, and economic status. Because the topic of intersectional identity is relatively new and exploration of the topic is in a nascent stage, the expertise required to support content review specific to a given intersectional identity may not exist at a level sufficient to form a review committee. Moreover, while considerable effort has been invested in developing item authoring guidelines and item review procedures that address issues specific to gender, racially stratified identity, economic status, accessibility, English language development, and other forms of potential bias, similar work focused on intersectional issues has not yet been conducted. If the field is to adopt an intersectional lens to the consideration of measurement bias, this work is requisite.

One approach to conducting this work is to apply an intersectional approach to each year's operational data to identify sub-groups for whom larger numbers of test items are flagged and to then focus attention on identifying possible causes of differential functioning for these groups. Although sufficient expertise may not yet exist to identify reasons for differential functioning, analysis over a period of time will support the development of this specialized body of knowledge. As an example, similar attention focused on science items administered to students developing English language proficiency unveiled several factors that contribute to lower than expected performance and helped establish item development guidelines for this sub-group of test takers (Noble et al., 2014a, 2014b).

A second issue specific to the review of items raised by the intersectional approach focuses on rethinking the criteria employed to forward an item for review. As noted previously, review of flagged items often fails to identify a content-specific cause of differential functioning. In such cases, the item is typically forwarded for operational use despite its differential functioning. If the findings presented here generalize to other tests, the number of items forwarded for review will increase substantially. This increase in review produces an increased cost for test development. From a practical perspective one might ask, if most reviewed items fail to identify a cause of differential functioning and are used operationally, is this increased cost worth it? To control unnecessary costs, one might then consider modifying review criteria such that only items flagged for a minimum number of subgroups (e.g., three or more) are forwarded for review. From an ethical perspective, however, one might argue that a test developer has a responsibility to take all reasonable steps to reduce measurement bias and thus is obligated to review any item flagged for any given focal group.

To date, the tradeoff between practicality and ethical responsibility has not been considered when using the traditional approach to group formation, in large part because the flagging of items of review is relatively uncommon. If the field is to adopt an intersectional approach, further consideration regarding this trade-off is warranted. In so doing, we encourage careful consideration of the concept of justice and advocate that the field adopt a conception aligned with Rawls (1971/91) theory of Justice as Fairness rather than the utilitarian (Sidgwick, 1907) theory that dominates our nation's social-economic-political structure. Whereas the utilitarian view allows net benefit to occur despite harm to some, a Justice as Fairness perspective requires all to benefit (although not necessarily equally). We suggest the current practice of retaining items that show statistical differential functioning but content review fails to identify a reason for such functioning is a utilitarian approach. Any modification to criteria that triggers content review that establishes a minimum threshold for the number of groups flagged would further increase utility by limiting the number of items that require replacement which decreases test development costs. Cost savings benefit testing programs and tax payers, but may produce harm for those students who form the focal group if the criteria for reviewing an item fails to examine an item that is actually biased. From a Justice as Fairness frame, one might hold that any item flagged for any focal group should be forwarded for review to help ensure (or at least minimize) measurement bias for all. One might further shift the rules governing the removal of an item from requiring a review panel to identify a construct-irrelevant factor that produces bias to a panel providing evidence that any difference in functioning of the item is construct-relevant. Robust consideration of the tradeoff between utility and Justice as Fairness are requisite when making modifications to review criteria to accommodate an intersectional approach to DIF analyses.

## Limitations

It is important to emphasize that this study focused only on the flagging of items as either suspicious or likely based on criteria established for the standardized D method. In an operational DIF analysis, flagged items that meet a given threshold are forwarded for review to identify construct-irrelevant factors that may cause the item to perform differentially between the flagged focal group(s) and the reference group. The study presented here was not able to perform this follow-up analysis and instead focused only on the impact that the method of defining groups had on the flagging of items. For this reason findings from this study should not be used to make interpretations about bias for or against any sub-groups of students.

The study also focused on only one of several tests administered by a state assessment program in a given year. Similarly, this study used only one of four commonly employed methods for examining DIF. Further analyses of other tests and other DIF detection methods are needed to determine if the findings presented here generalize across subject areas, grade levels, and methods.

A final limitation of this study is its focus on the intersection of three demographic characteristics, namely gender, racial stratification, and economic status. There are other demographic characteristics and administrative conditions that are often of concern in a DIF analyses, including first language, special education status, and use of different types of digital tools and/or accessibility supports. However, before adding additional demographic characteristics and/or administrative conditions to form more nuanced groupings, it is important to establish a theory as to why the various identities and conditions comprising a grouping might intersect to produce disparate impact. In U.S. society, it is well established that people who are White, people who are male, and people with greater financial resources are advantaged. It is also understood that one's gender, racially stratified identity, and economic status interact to impact the extent to which a given individual is advantaged or disadvantaged within the U.S. socio-political-economic system. Adding other demographic characteristics and/or administrative conditions, such as the locale in which one lives or type of hardware used during test administration (e.g., laptop computer, desktop computer, or tablet) require theory as to why and how they might interact with other demographic variables to impact advantage/disadvantage before applying them to form an intersectional identity. Clearly expanding the demographic characteristics and/or administrative conditions that define an intersectional group will require careful consideration, both in terms of establishing a sound rationale for the group's relationship to potential item bias and the impact its inclusion will have on sample size requirements.

Despite these limitations, this study provides preliminary evidence that the method used to define groups that are the focus of differential item functioning, and sources of potential bias in test scores, does matter. The intersectional approach employed here resulted in a substantially higher percentage of flagged items and indicated that, for some intersectional groups, a substantial number of items may contribute to bias in test scores. Based on these initial findings, further efforts to compare the traditional and intersectional methods for other subject areas, grade levels, and testing programs is advised. Research should also examine whether these findings hold across different methods for examining DIF.

In making these suggestions, we anticipate that re-orientating DIF analyses to incorporate both intersectional identity and a Justice as Fairness frame will not sit well with some readers. We acknowledge this re-orientation runs counter to current practice and may seem impractical to implement operationally. Such reactions are understandable and consistent with those that occurred in response to early advocacy for test accommodations for students with disabilities, providing flexible accessibility options for all students, and adopting interoperability standards for digital item content. While each of these concepts presented practical challenges to test development and administration, each has now been implemented at scale and has become common practice in the testing industry. Although more research is needed to explore and address the challenges presented by an intersectional approach to DIF analyses, we believe the same potential holds for examining item bias through the compound lens of intersectional identity and Justice as Fairness.

## References

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

Baker, C. A. (2019). A QuantCrit Approach. *Journal of Underrepresented & Minority Progress*, *3*(1), 1-22.

Belzak, W.C.M. (2020) Testing Differential Item Functioning in Small Samples, *Multivariate Behavioral Research*, 55:5, 722-747, DOI: 10.1080/00273171.2019.1671162

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 57, 289–300.

Chapa, J. & Schink, W. (2006). California Community Colleges: Help or Hindrance to Latinos in the Higher Education Pipeline?. *New Directions for Community Colleges*, *133*, 41-5.

Crenshaw, K. (1991). Mapping the margins: Identity politics, intersectionality, and violence against women. *Stanford Law Review*, *43*(6), 1241-1299.

Cuevas, M. & Cervantes, V. H. (2012). Differential item functioning detection with logistic regression. *Mathématiques et sciences humaines. Mathematics and social sciences*, (199), 45-59.

Denzin, N. K. (2017). Critical qualitative inquiry. *Qualitative inquiry*, *23*(1), 8-16.

Dorans, N. J. & Kulick, E. (1986). Demonstrating the Utility of the Standardization Approach to Assessing Unexpected Differential Item Performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, 23(4), 355-368.

Dorans, N. J. & Holland, P. W. (1992). DIF Detection and Description: Mantel‐Haenszel and Standardization 1, 2. *ETS Research Report Series*, *1992*(1), i-4.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement*, *29*(4), 309-319.

Fidalgo, A. M., Ferreres, D., & Muniz, J. (2004). Utility of the Mantel-Haenszel procedure for detecting differential item functioning in small samples. *Educational and Psychological Measurement*, *64*(6), 925-936.

Garcia, N. M., Lopez, N., & Velez, V. N. (2018). QuantCrit: Rectifying quantitative methods through critical race theory. *Race Ethnicity and Education*, 21(2), 149-157, DOI: 10.1080/13613324.2017.1377675

Gillborn, D. (2010). The colour of numbers: surveys, statistics and deficit‐thinking about race and class. *Journal of Education Policy*, *25*(2), 253-276.

Gillborn, D., Warmington, P., & Demack, S. (2018). QuantCrit: education, policy, 'Big Data', and principles for a critical race theory of statistics. *Race Ethnicity and Education*, *21*(2), 158-179.

Hancock, A. M. (2013). "Empirical Intersectionality: A Tale of Two Approaches." UC Irvine L. Rev. 3: 259.

Holland, P. W. (2008). Causation and race. Zuberi, T., & Bonilla-Silva, E. (Eds.). (2008). *White logic, white methods: Racism and methodology,* 93-109. Rowman & Littlefield Publishers

Holland, P. W. & Wainer, H. (1993). *Differential Item Functioning*. Lawrence Erlbaum Associates.

Horn, L. J. (1997). *Confronting the Odds: Students At Risk and the Pipeline to Higher Education. Statistical Analysis Report*. US Government Printing Office, Superintendent of Documents, Mail Stop: SSOP, Washington, DC 20402-9328.

King, J. E. (2000). Gender Equity in Higher Education: Are Male Students at a Disadvantage?.

LaVeist, T. A. (1994). Beyond Dummy Variables and Sample Selection: What Health Services Researchers Ought to Know about Race as a Variable. *Health Services Research* 29(1): 1–16.

López, N., Erwin, C., Binder, M., & Chavez, M. J. (2018). Making the invisible visible: Advancing quantitative methods in higher education using critical race theory and intersectionality. *Race Ethnicity and Education*, *21*(2), 180-207.

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

Massachusetts Department of Elementary and Secondary Education. (2018) *2018 Next-Generation MCAS and MCAS-Alt Technical Report*. Malden, MA.

Mazon, M. R. & Ross, H. (1990). Minorities in the higher education pipeline: a critical review. *The Western Journal of Black Studies*, *14*(3), 159.

McCall, L. (2001). *Complex inequality: Gender, class, and race in the new economy*. Psychology Press.

McCall, L. (2005). "The Complexity of Intersectionality." *Signs* 30(3): 1771–180.

Museus, S. D. and K. A. Griffin. (2011). Mapping the Margins in Higher Education: On the Promise of Intersectionality Frameworks in Research and Discourse. *New Directions for Institutional Research,* 2011(151): 5–13.

Noble, T., Kachchaf, R., Rosebery, A., Warren, B., O'Connor, M. C., & Wang, Y. (2014a). Do linguistic features of science test items prevent English language learners from demonstrating their knowledge? Paper presented at the annual meeting of the National Association of Research on Science Teaching, Pittsburgh, PA.

Noble, T., Rosebery, A., Suarez, C., Warren, B., & O'Connor, M. C. (2014b). Science assessments and English language learners: Validity evidence based on response processes. *Applied Measurement in Education*, *27*(4), 248-260.

Oklahoma State Department of Education. (2019). *Oklahoma School Testing Program/College and Career-Readiness Assessment Grades 2-8, 11*. Oklahoma City, OK.

Oliver, M. L. & Shapiro, T. M. (1989). Race and wealth. *The Review of Black Political Economy*, *17*(4), 5-25.

Oliver, M. L. & Shapiro, T. M. (2001). Wealth and racial stratification. *America Becoming*, *2*, 222-251.

Oliver, M. L. Shapiro, T. M., & Shapiro, T. (2006). *Black wealth, white wealth: A new perspective on racial inequality*. Taylor & Francis.

Osterlind, S. J. & Everson, H. T. (2009). *Differential item functioning* (Vol. 161). Sage Publications.

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.

Rawls, J. (1971/1991). *A theory of justice*. Harvard university press.

Rothstein, R. (2017). *The color of law: A forgotten history of how our government segregated America*. Liveright Publishing.

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement*, 143-152.

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

Sidgwick, H. (1907) *The Methods of Ethics*, Seventh Edition. Macmillan.

Somes, G. W. (1986). The generalized Mantel–Haenszel statistic. *The American Statistician*, *40*(2), 106-108.

Stage, F. K. & Wells, R. S. (2014). Critical quantitative inquiry in context. *New Directions for Institutional Research*, *2013*(158), 1-7.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, *27*(4), 361-370.

Zuberi, T. (2001). *Thicker than Blood: How Racial Statistics Lie*. University of Minnesota.

Zwick, R. (2012). *A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement*. ETS Research Report RR-12-08.

**Citation**:
Russell, M., & Kaplan, L. (2021). An Intersectional Approach to Differential Item Functioning: Reflecting Configurations of Inequality. *Practical Assessment, Research & Evaluation*, 26(21). Available online: https://scholarwoarks.umass.edu/pare/vol26/iss1/21/

**Corresponding Author**

Michael Russell
Lynch School of Education and Human Development
Boston College
Chestnut Hill, MA USA

Email: russellmh [at] bc.edu