Investigating Assessment Conditions Potentially Associated with Differential Item Functioning

Qiao Lin and Jeffrey Steedle, PhD

Background

Large-scale assessment programs commonly field test newly developed items to evaluate their psychometric properties and determine whether the items are suitable for future operational administration. In field test item analyses, items may be flagged for a variety of reasons, including being too easy or too difficult, exhibiting low discrimination (i.e., providing little information about which examinees have lower or higher ability), or having distractors that are chosen at an unexpectedly high rate. In addition, field-tested items are commonly examined for statistical evidence of differential item functioning (DIF) to help prevent administering items that could unfairly disadvantage certain examinee groups.

Assessment programs develop new test forms on a regular basis, and certain programs demand many new forms because of frequent administrations and multiple testing modes. The ACT® test, for example, needs forms for national Saturday test dates, state and district administrations, and international administrations. With such high demand for new forms, losing items during field test analyses poses significant challenges, especially for passage-based tests where a whole passage can be lost if too many items are rejected.

ACT items are developed in a manner that promotes accessibility, and all items are reviewed for bias and sensitivity issues prior to field testing. Yet, in the course of field test analyses, some items will be flagged for DIF. Those items are reviewed by a panel of content experts that is diverse in terms of gender and race/ethnicity to identify sources of potential item bias (e.g., the item includes content that would be familiar to one examinee group but not another). Practically none of these reviews identify reasons that items might be biased. This can lead to difficult conversations between test development teams and psychometricians. For example, mathematics items containing only numbers and symbols are sometimes flagged for DIF. Sometimes an item is flagged for DIF even when numerous similar items are not. The test development team wants to know, “What is wrong with these items?” Meanwhile, psychometricians are left to wonder, “Are these items truly biased, or is the observed DIF a statistical artifact or Type-I error (i.e., false positive)?”

This challenging situation was the motivation behind this study, which was designed to identify assessment conditions potentially associated with DIF as indicated by the Mantel-Haenszel procedure (Holland & Thayer, 1988). In this study, analyses of empirical data from ACT field testing revealed that easier items were more likely to exhibit DIF favoring females, and harder items were more likely to favor males. Highly discriminating items were more likely to favor White examinees, whereas less discriminating items were more likely to favor minority examinees. A subsequent simulation study examined how test length, item difficulty, item discrimination, examinee ability, and examinee sample size were related to statistical evidence of DIF. Results indicated that none of those factors nor their interactions caused spurious DIF flags at an unexpected rate. Thus, from a statistical perspective, the Mantel-Haenszel DIF detection procedure behaved as expected. Consequently, when the percentage of items flagged for DIF is low (e.g., near 5%), a sizeable proportion of those flags could represent Type-I errors. False positive DIF flags could be reduced by increasing flagging thresholds, but this would also reduce the number of item accurately flagged for DIF. Overall, these results provide support for the practice of reviewing items flagged for slight to moderate DIF and approving them for administration when content reviews identify no cause for bias.

Prior Research

In general, DIF analyses attempt to identify potentially biased test items according to whether one examinee group performs unusually well or poorly on an item compared to another examinee group when controlling for overall ability. Conventionally, DIF analyses compare the “reference” group (e.g., White examinees, male examinees, native English speakers, etc.) to the “focal” group of concern (e.g., minority examinees, female examinees, English language learners, etc.). Previous studies have examined the relationship between DIF and other properties of items. Freedle and Kostin (1990), for example, examined the relationship between item difficulty and a DIF index (the standardized difference between p-values) across four types of verbal items from the GRE and SAT. There was a positive correlation indicating that Black examinees performed differentially better on more difficult items and differentially worse on the easier items compared to White examinees. Freedle, Kostin, and Schwartz (1987) conducted a think-aloud study to account for this pattern in DIF analysis results. Their results suggested that Black examinees were more likely to use certain strategies (e.g., indirect induction and word associations) that were related to probability of responding correctly to certain item types. Later, Freedle and Kostin (1997) investigated factors related to DIF between Black and White examinees on verbal items. They concluded that results were consistent with the “cultural familiarity hypothesis.” Namely, easier items tended to deal with concepts differentially less familiar to minority examinees, and more difficult items tended to deal with concepts differentially more familiar to minority examinees.

Despite criticism of Freedle’s choices of data sets and methodologies, researchers have replicated the relationship between item difficulty and DIF observed by Freedle and his colleagues, and they have studied whether this relationship could be an artifact of the statistical techniques used in DIF analyses. For example, Schernaum and Goldstein (2008) found the same pattern in results using IRT-based DIF methods. Similarly, Santelices and Wilson (2012) used several IRT models to explore the

relationship between item difficulty and DIF, and their results also showed that easier items tended to favor Black examinees, and harder items tended to favor White examinees. They concluded that results of earlier research were not statistical artifacts related to DIF analysis methods or choice of data set.

Besides item difficulty, the relationship between DIF and other characteristics of items have been investigated using simulation studies. Mazor, Clauser, and Hambleton (1992) conducted a simulation study to examine the sample size required to detect varying types and levels of DIF. Results indicated that the DIF detection rates decreased when sample size was smaller. The authors concluded that a sample size of 200 is acceptable if only items with severe DIF are of concern. A simulation study conducted by Fidalgo, Mellenbergh, and Muñiz (2000) examined the effect of different numbers of DIF items, test lengths, and purification methods (i.e., removing flagged items from DIF analyses) on robustness and power of Mantel-Haenszel (MH) DIF procedures. The authors observed that power increased slightly with test length and that Type-I error rates tended to be greater when a larger number of items exhibited DIF.

The current study adds to the research literature first by examining relationships between the psychometric properties of items (difficulty and discrimination) and MH DIF results (White vs. minority and male vs. female) using data from ACT item field testing. Second, a simulation study addresses the research question, “What assessment conditions are associated with spurious DIF flags?” Those conditions included varying item difficulty, item discrimination, test length, focal group ability, and focal group sample size. Results of the study have practical implications because, if certain conditions are associated with relatively low or high Type-I error rates, that could influence the interpretation of certain DIF analysis results.

Method

Data

When students take the ACT test, their test booklets include English, reading, math, and science sections followed by a short booklet of field test items from one of the four content areas. Data from the field test booklets are analyzed by ACT psychometricians and reviewed by content staff to determine whether items are suitable for use on future ACT forms. ACT field testing is designed to collect approximately 1,000 responses to each item, which allows for DIF analyses comparing male and female examinees as well as White and minority examinees with a minimum sample size of 200 per group and 500 total. The minority group included students from the following racial/ethnic groups: Black/African American (approximately 43%), American Indian/Alaska Native (2%), Hispanic/Latino (37%), Asian (17%), and Native Hawaiian/Other Pacific Islander (1%). These groups are combined in field test analyses because any individual group would not meet minimum sample size requirements. The data analyzed for this study came from the field test item analyses conducted after the ACT administrations in September 2019, October 2019, December 2019, and February 2020. The data included 2,451 English items, 3,361 math items, 1,687 reading items, and 2,705 science items.

Note that field test items failing to meet certain statistical criteria are not eligible for use on future ACT forms. This includes meeting acceptable ranges for item difficulty and discrimination, exhibiting acceptable response distributions, and passing item reviews triggered by DIF analyses. Specifically, items exhibiting moderate to large DIF are not eligible, but items exhibiting slight to moderate DIF may be eligible pending review by a panel of content experts. Eligibility depends on whether the panel identifies any reason an item could be biased.

The simulation study used 3PL IRT model parameters estimated using data from the February 2020 ACT equating study. Each year, ACT administers new forms to a sample of examinees for the purpose of equating those forms to the ACT 1–36 score scale. The equating sample is selected to be representative of the ACT examinee population in terms of ability, with a sample size of 2,000 or more examinees taking each new form. After equating, 3PL item parameters (*a, b*, and *c*) are estimated for each item, and those parameters are transformed to a common scale using the Stocking-Lord method (Stocking & Lord, 1983). The data for this study included items from 16 new forms, each with 75 English items, 60 math items, 40 reading items, and 40 science items.

DIF Detection

Holland and Thayer (1988) introduced the Mantel-Haenszel (1959) method to detect DIF between matched groups of examinees. With the selected matching criteria (e.g., total raw scores), the data can be arranged into a series of 2 × 2 frequency tables with groups in the rows and item scores in the columns. One table is generated for each matched set of examines in the reference group (*R*) and focal group (*F*). Table 1 represents the *j*th matched set (or the jth "stratum") for item *i*.

Table 1. 2 × 2 Frequency Table

Group	Item Score		Total
	1	0	
R	A_j	B_j	n_{Rj}
F	C_j	D_j	n_{Fj}
Total	m_{1j}	m_{0j}	T_j

Given the data in the 2 × 2 tables, Mantel and Haenszel developed a chi-squared test with null and alternative hypotheses:

$$H_0: \frac{p_{Rj}}{q_{Rj}} = \frac{p_{Fj}}{q_{Fj}}$$

$$H_1: \frac{p_{Rj}}{q_{Rj}} = \alpha \frac{p_{Fj}}{q_{Fj}}$$

for all strata $j = 1, 2, \ldots, K$, where $\frac{p_{Rj}}{q_{Rj}}$ represents the odds of correct response in the reference group and $\frac{p_{Fj}}{q_{Fj}}$ represents the odds of correct response in the focal group for the j^{th} stratum (p is probability of correct response, q is probability of incorrect response). Under H_1, the parameter α is the odds ratio $\frac{p_{Rj}q_{Fj}}{p_{Fj}q_{Rj}}$ for all strata.

The Mantel-Haenszel chi-squared test statistic is defined as

$$MH\ CHISQ = \frac{(|\sum_j A_j - \sum_j E(A_j)| - \frac{1}{2})^2}{\sum_j Var(A_j)} \quad (1)$$

As indicated by Table 1, A_j represents the number of examinees who gave the correct response in the reference group. Note that the subtraction of 1/2 is a continuity correction to improve the approximation of the observed significance levels using the chi-squared table. Under H_0, *MH CHISQ* has a chi-squared distribution with one degree of freedom. The expected value and the variance of A_j are defined as follows:

$$E(A_j) = \frac{n_{Rj} m_{1j}}{T_j} \quad (2)$$

$$Var(A_j) = \frac{n_{Rj} n_{Fj} m_{1j} m_{0j}}{T_j^2 (T_j - 1)} \quad (3)$$

where n_{Rj} and n_{Fj} are the numbers of examinees in the reference and focal groups, and m_{1j} and m_{0j} are the number of examinees who answered the item correctly and incorrectly, respectively, regardless of groups.

In addition, Mantel and Haenszel defined an estimate of the common odds ratio α as

$$\hat{\alpha}_{MH} = \frac{\sum_j A_j D_j / T_j}{\sum_j B_j C_j / T_j} \quad (4)$$

where B_j and D_j represent the numbers of examinees who gave incorrect responses in the reference group and focal groups, respectively, and T_j is the total number of examinees. This estimator indicates the degree to which the data depart from H_0. When α=1, there is no DIF. That is, the odds of responding correctly are the same in the reference and focal groups.

To transform $\hat{\alpha}_{MH}$ to a symmetric scale, Holland and Thayer (1988) proposed the *MH D-DIF* index defined as

$$MH\ DDIF = -2.35 \ln \hat{\alpha}_{MH} \quad (5)$$

MH D-DIF (or Δ_{MH}) is scaled to the ETS delta scale of item difficulty. A negative value of *MH D-DIF* indicates DIF favoring the reference group; a positive value indicates DIF favoring the focal group.

For a dichotomous item, classification rules were developed to classify the item into three categories (A, B, and C) according to the severity of the DIF (Dorans & Holland, 1993; Zieky, 1993). Furthermore, items labeled B and C are distinguished by their signs: B+ and C+ items exhibit DIF favoring the focal group, whereas B- and C- items exhibit DIF favoring the reference group. The conventional flagging rules to classify items are summarized as follows:

Table 2. Mantel-Haenszel DIF Classification Rules

DIF Flag	Flagging Rules
A: nonsignificant DIF	*MH CHISQ* is not significant at the .05 level (i.e., *MH CHISQ* ≤ 3.84) or |MH D-DIF|<1
B: slight to moderate DIF	*MH CHISQ* > 3.84 and |*MH D-DIF*| ≥ 1
C: moderate to large DIF	(|*MH D-DIF*|-1)/SE(*MH D-DIF*) > 1.645 (i.e., *MH D-DIF* is significantly greater than 1 in absolute value at the .05 level) and |*MH D-DIF*|≥ 1.5

Analysis

Empirical Analysis

A descriptive analysis was conducted to summarize the relationships between DIF and the psychometric properties of ACT field test items. As part of field test analyses, measures of item difficulty (item proportion correct or "p-value") and discrimination (point-biserial correlation) were calculated, and all items were analyzed for DIF using the MH approach (male vs. female and White vs. minority). Examinees' raw scores on the operational items were used to divide them into 10 strata. The available data files provided only the DIF classifications (C-, B-, A, B+, or C+), not statistics such as *MH CHISQ* or *MH D-DIF*. For this study, the associations between difficulty and DIF and between discrimination and DIF were examined for the four ACT sections: English, math, reading, and science.

Simulation Study

The simulation study was designed to investigate the relationship between DIF and assessment conditions. Considering prior research, several conditions were included in this study: test length, item difficulty, item discrimination, focal group sample size, and focal group ability (Table 3). The simulation study was repeated with item parameters from two ACT sections: English and math.

Three test lengths were used in the simulation: 25 items, 50 items, and 75 items. Likewise, there were three item difficulty conditions. The middle difficulty condition was generated by randomly sampling from the available items, which had a mean difficulty (*b*) parameter of -0.18 and standard deviation of 0.86 for English and mean of 0.29 and standard deviation of 1.27 for math. For the easy and difficult conditions, a similar random sampling procedure was implemented, but the *b* parameters were adjusted by -1 for the easy condition and +1 for the difficult condition. Similarly, low, middle, and high levels of item discrimination were studied. This was achieved by dividing items into thirds according to their IRT discrimination (*a*) parameter estimates.

The other varying conditions of the simulation were focal group sample size and focal group ability. For the reference group, examinee ability (θ) was generated from a normal distribution with a mean of 0 and standard deviation of 1. The sample size of the reference group was always 500, but the focal group varied (100, 200, and 300). The mean ability of the focal group took on three values: -1 for the low ability condition, 0 for the middle ability condition (same as the reference group), and 1 for the high ability condition (always with a standard deviation of 1).

Prior studies have shown that Type-I error rates can be inflated when a larger number of items exhibit DIF (e.g., Fidalgo et al., 2000), and it is common for an assessment to have some items exhibiting DIF. For these reasons, the simulation study was repeated under two DIF conditions: zero DIF items and 10% simulated DIF items. Ten percent was chosen because nearly 10% of field test items analyzed were flagged for DIF. In the zero DIF condition, none of the simulated items truly exhibited DIF, so any DIF flags were false positives. In the simulated DIF condition, 10% of items were randomly selected to exhibit DIF by adjusting their *b* parameters by 0.5 in a randomly selected group (reference or focal). In that analysis, it was possible to calculate the detection rate (true-positive rate or "sensitivity") as the proportion of true DIF items flagged by the MH procedure.

Each simulation study included a total of 243 conditions (3 test length conditions × 3 item difficulty conditions × 3 item discrimination conditions × 3 focal group sample size conditions × 3 focal group ability conditions). Each of the 243 simulation conditions was replicated 100 times. That study design was repeated a total of four times (2 subject tests × 2 DIF conditions). A 3PL IRT model was used to simulate examinees' 0/1 scores on items selected from the available item pool. In each replication, four outcomes were recorded: the proportion of items correctly flagged for DIF (in the simulated DIF condition), the proportion of items incorrectly flagged for DIF, the mean of *MH D-DIF*, and the proportion of items with no DIF having a statistically significant *MH CHISQ* value (i.e., *MH CHISQ* Type-I error rate).

Table 3. Simulation Study Conditions

Condition	Value
Test Length	25 items
	50 items
	75 items
Item Difficulty	Easy test (English: mean b = –1.18, SD = 0.86; math: mean b = –0.71, SD = 0.86)
	Medium test (English: mean b = –0.18, SD = 0.86; math: mean b = 0.29, SD = 1.27)
	Hard test (English: mean b = 0.82, SD = 0.86; math: mean b = 1.29, SD = 1.27)
Item Discrimination	Low item discrimination (English: $a \leq 0.66$; math: $a \leq 0.68$)
	Middle item discrimination (English: $0.66 < a < 0.87$; math: $0.68 < a < 0.90$)
	High item discrimination (English: $0.87 \leq a$; math: $0.90 \leq a$)
Focal Sample Size	Focal group 100, reference group 500
	Focal group 200, reference group 500
	Focal group 300, reference group 500
Focal Ability	Focal group ~ N(-1, 1), reference group ~ N(0, 1)
	Focal group ~ N(0, 1), reference group ~ N(0, 1)
	Focal group ~ N(1, 1), reference group ~ N(0, 1)

Results

Descriptive Analysis

Tables 4 and 5 show descriptive statistics for ACT field test items that were analyzed for gender DIF and race/ethnicity DIF, respectively. Specifically, the tables show descriptive statistics for proportion correct (item difficulty) and point-biserial correlation (item discrimination) of field test items across four subjects by MH DIF classification (A, B, or C). Approximately 94% of items had a DIF classification of "A" (nonsignificant DIF) in the gender or race/ethnicity analysis. In the gender DIF analysis, there was a tendency for math items to favor males (e.g., 116 B- items vs. 54 B+ items). To a lesser degree, math items were also more likely to favor White examinees in the race/ethnicity analysis. Note that 25% of items flagged for DIF "failed" data review for other reasons (e.g., item difficulty or discrimination fell outside the acceptable range).

Figure 1 illustrates the associations between DIF classifications and item statistics for the gender DIF analysis. Mean p-values tended to be lower for items favoring males (C- and B-) and higher for items favoring females (B+ and C+). That is, more difficult items tended to favor males, and easier items tended to favor females. This trend was apparent across all four subject areas. There was a weaker association between item discrimination and DIF classifications, wherein more discriminating items tended to favor males, and less discriminating items tended to favor females. These results were expected because of the negative correlation between p-values and point-biserial correlations, especially for the English test (r = -0.92) and the science test (r = -0.74). That is, easier items tended to be less discriminating, and harder items tended to be more discriminating.

In the race/ethnicity analysis, there was not a clear association between difficulty and DIF classification except on the science test, where more difficult items (lower p-values) were more likely to favor minority examinees, and easier items (higher p-values) were more likely to favor minority examinees (Figure 2). Note that the means for C- and C+ items were unreliable due to small sample sizes on the science test (1–2 items), but the trend was still apparent on the B-, A, and B+ items. On the English, math, and reading tests, easier items were more likely to be classified as C- or C+. On the English, reading, and science tests, items with higher mean discrimination were more likely to favor White examinees, and items with lower mean discrimination were more likely to favor minority examinees.

Table 4. Descriptive Statistics for Field Test Item Difficulty and Discrimination by Gender DIF Classification

				Proportion Correct				Point-Biserial Correlation			
Subject	MH Flag	N	%	Mean	SD	Min	Max	Mean	SD	Min	Max
	C-	16	0.7%	0.62	0.19	0.34	0.92	0.36	0.08	0.21	0.49
	B-	87	3.5%	0.62	0.20	0.20	0.96	0.34	0.09	0.14	0.58
English	A	2261	92.2%	0.67	0.18	0.07	0.99	0.32	0.10	-0.08	0.59
	B+	73	3.0%	0.75	0.16	0.19	0.98	0.33	0.08	0.06	0.50
	C+	14	0.6%	0.84	0.18	0.39	0.99	0.28	0.09	0.16	0.47
	C-	25	0.7%	0.32	0.21	0.06	0.83	0.33	0.06	0.17	0.43
	B-	116	3.5%	0.39	0.21	0.07	0.93	0.33	0.09	0.08	0.49
Math	A	3152	93.8%	0.42	0.21	0.00	0.99	0.26	0.12	-0.23	0.57
	B+	54	1.6%	0.61	0.25	0.14	0.99	0.31	0.08	0.15	0.45
	C+	14	0.4%	0.70	0.29	0.00	0.99	0.29	0.12	0.08	0.44
	C-	7	0.4%	0.47	0.21	0.20	0.86	0.27	0.06	0.16	0.33
	B-	32	1.9%	0.64	0.13	0.41	0.85	0.33	0.08	0.18	0.46
Reading	A	1611	95.5%	0.62	0.16	0.09	0.99	0.27	0.09	-0.11	0.49
	B+	33	2.0%	0.79	0.12	0.52	0.99	0.28	0.06	0.16	0.39
	C+	4	0.2%	0.83	0.17	0.59	0.95	0.24	0.03	0.22	0.28
	C-	18	0.7%	0.38	0.18	0.07	0.76	0.30	0.12	-0.02	0.43
	B-	50	1.8%	0.53	0.21	0.08	0.94	0.30	0.11	-0.01	0.48
Science	A	2569	93.4%	0.58	0.19	0.04	0.99	0.25	0.10	-0.14	0.52
	B+	62	2.3%	0.78	0.18	0.18	0.99	0.28	0.09	-0.08	0.43
	C+	6	0.2%	0.92	0.09	0.80	0.99	0.23	0.05	0.18	0.30

Table 5. Descriptive Statistics for Field Test Item Difficulty and Discrimination by Race/Ethnicity DIF Classification

				Proportion Correct				Point-Biserial Correlation			
Subject	MH Flag	N	%	Mean	SD	Min	Max	Mean	SD	Min	Max
	C-	20	0.8%	0.78	0.19	0.20	0.95	0.38	0.05	0.26	0.50
	B-	69	2.8%	0.75	0.16	0.24	0.98	0.36	0.08	0.09	0.55
English	A	2301	93.9%	0.67	0.18	0.07	0.99	0.32	0.10	-0.07	0.59
	B+	54	2.2%	0.70	0.19	0.12	0.96	0.29	0.09	0.04	0.43
	C+	7	0.3%	0.76	0.24	0.41	0.97	0.18	0.14	-0.08	0.35
	C-	32	1.0%	0.61	0.30	0.01	0.98	0.37	0.08	0.11	0.52
	B-	129	3.8%	0.62	0.24	0.14	0.99	0.37	0.08	0.17	0.51
Math	A	3089	91.9%	0.41	0.20	0.00	0.99	0.26	0.12	-0.11	0.57
	B+	87	2.6%	0.39	0.21	0.03	0.80	0.22	0.14	-0.23	0.48
	C+	24	0.7%	0.66	0.27	0.15	0.99	0.33	0.14	-0.13	0.49
	C-	10	0.6%	0.69	0.25	0.19	0.99	0.36	0.08	0.19	0.47
	B-	34	2.0%	0.70	0.14	0.40	0.90	0.34	0.06	0.20	0.47
Reading	A	1609	95.4%	0.62	0.16	0.09	0.99	0.27	0.09	-0.11	0.49
	B+	33	2.0%	0.68	0.15	0.35	0.94	0.25	0.08	0.06	0.39
	C+	1	0.1%	0.73	--	0.73	0.73	0.17	--	0.17	0.17
	C-	2	0.1%	0.86	0.16	0.75	0.97	0.46	0.09	0.39	0.52
	B-	43	1.6%	0.65	0.20	0.22	0.98	0.35	0.07	0.21	0.51
Science	A	2626	95.5%	0.58	0.20	0.04	0.99	0.26	0.10	-0.14	0.50
	B+	33	1.2%	0.53	0.22	0.21	0.93	0.17	0.12	-0.08	0.38
	C+	1	0.0%	0.25	--	0.25	0.25	-0.08	--	-0.08	-0.08

Figure 1. Mean Item Difficulty and Discrimination by Gender DIF Classification

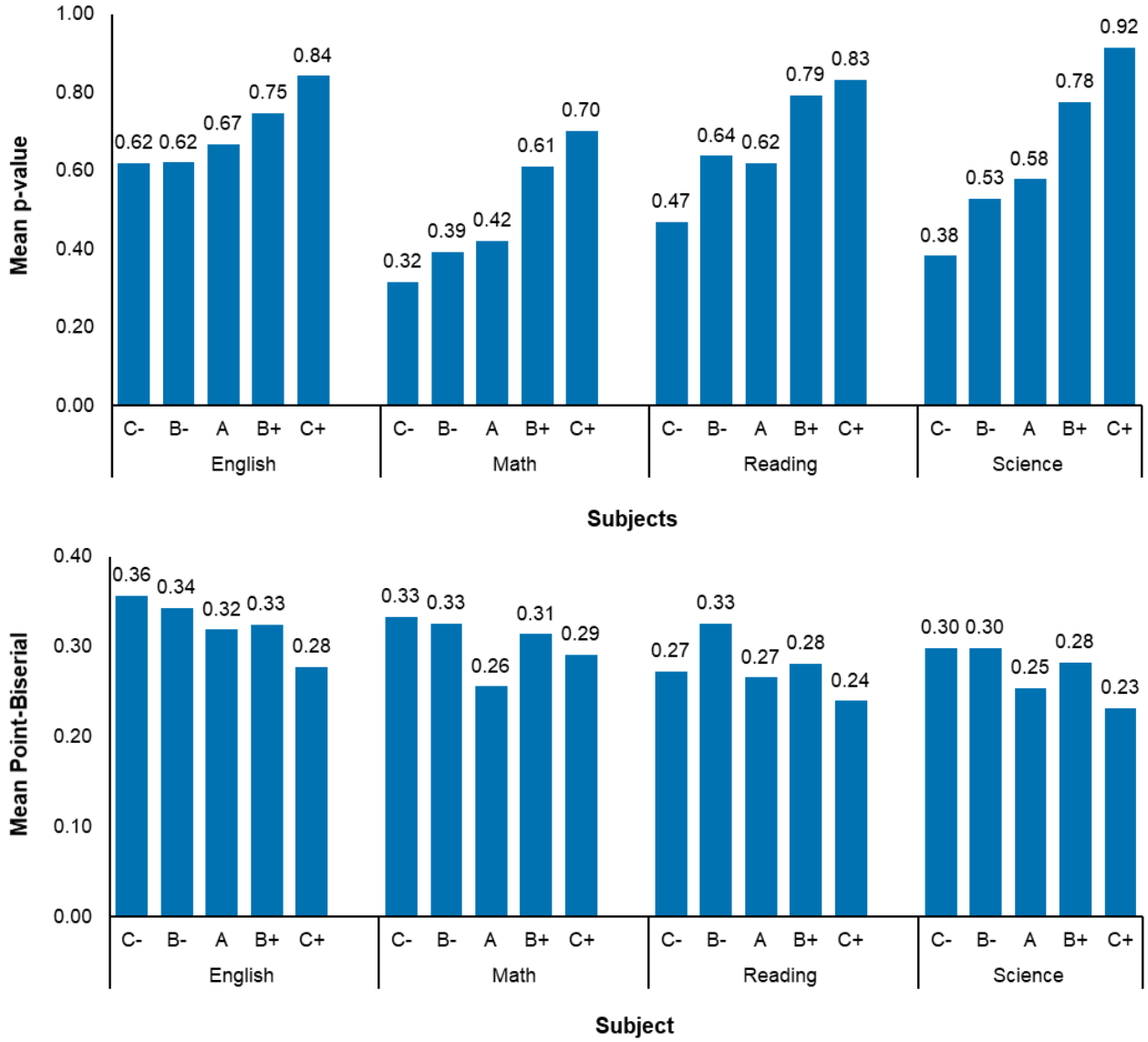
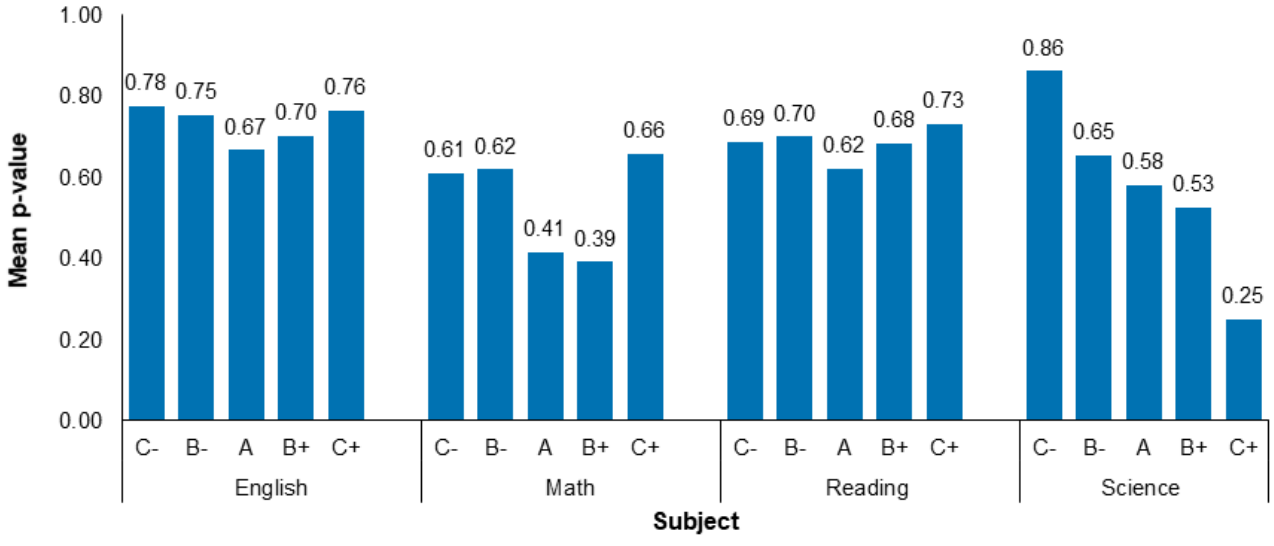
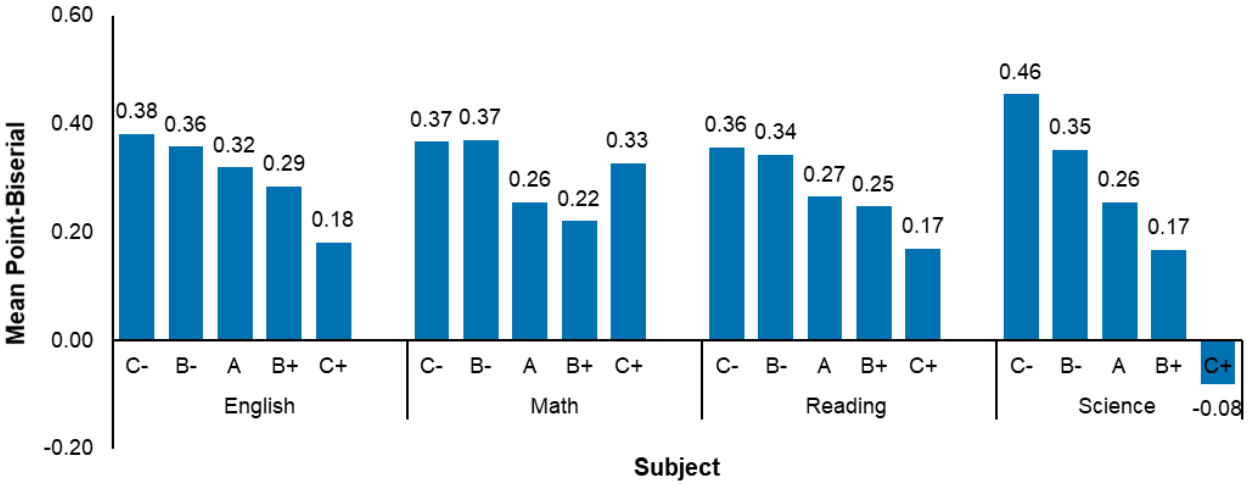


Figure 2. Mean Item Difficulty and Discrimination by Race/Ethnicity DIF Classification





Simulation Study

The simulation study identified main effects and interactions between assessment conditions on DIF results for the English and math tests. Table 6 illustrates possible main effects of various assessment conditions on the mean proportion of flagged DIF items, mean *MH D-DIF*, and mean *MH CHISQ* Type-I error rate across 100 replications for the English test when no items exhibited DIF. For example, from this table, it would be possible to observe how test length related to DIF results when averaging across all other assessment conditions. Overall, the mean proportion of items with B or C DIF classifications was low across all conditions (< .04) and changed little with differences in test length, test difficulty, test discrimination, focal group sample size, and focal group ability. Mean *MH D-DIF* increased slightly when the test had lower difficulty or higher focal group ability. The latter was expected to increase *MH D-DIF* because it would result in higher item performance for the focal group relative to the reference group. Type-I error rate increased slightly with higher test difficulty, larger focal group sample size, and lower focal group ability. Having larger sample sizes possibly led to more Type-I errors because chi-squared tests are known to be sensitive to sample size. In contrast, the proportion of items flagged for B or C DIF decreased slightly with greater sample size. All observed trends were very weak, and the overall rates of flagging items for DIF were at or below expected, which was .05 for the MH DIF approach.

Analysis of variance (ANOVA) was conducted to identify significant interaction effects between assessment conditions on the outcome variables. Omega squared (ω^2) provided an estimate of the effect size of the interactions. There was only one interaction with an ω^2 value notably different from zero: the two-way interaction between test difficulty and focal group ability ($\omega^2 = .07$). As shown in Table 7, the combination of low test difficulty and high focal group ability led to higher average *MH D-DIF*.

Table 8 shows descriptive statistics illustrating possible main effects in the analysis of math items. Trends in results from the math analyses were identical to the English analyses, though a slightly higher proportion of math items were flagged for DIF. Even the interaction analysis for math showed results similar to English (Table 9). Thus, the simulation methods used in this study appear to be robust to test content. Any differences in results could have been related differences in difficulty of the English and math tests.

The replication study with 10% simulated DIF items estimated the detection rate in each assessment condition and revealed the effects of having simulated DIF on the mean proportion of non-DIF items flagged for DIF, *MH D-DIF* for non-DIF items, and *MH CHISQ* Type-I error rate. The detection rate was the proportion of true DIF items flagged by the MH procedure (i.e., true positive rate). Table 10 shows that the detection rate ranged from 0.560 to 0.813. Thus, more than half of the true DIF items were identified by the MH procedure across the 100 replications of the 243 assessment conditions.

Note that patterns in results were the same for English and math. There was a weak main effect for test length wherein the detection rate was highest for tests with 50 items. Detection rates were lower for high difficulty tests compared to low and middle difficulty tests. The detection rate was higher when the focal sample size was 200 compared with 100 and 300. Discrimination exhibited one of the strongest associations with detection rate. Namely, when tests comprised items with higher discrimination, the detection rate was higher. The next strongest association was between focal ability and detection rate. Specifically, the detection rate was higher when the focal ability was low. Interactions between testing conditions were also examined as predictors of the detection rate, but ω^2 was close to zero in all cases.

Compared to results of simulation study with no simulated DIF items, the proportions of non-DIF items flagged for B or C DIF and the *MH CHISQ* Type-I error rates were very similar. The only notable change occurred on the *MH D-DIF* statistics for the low and high focal ability groups. When 10% of items had simulated DIF, mean *MH D-DIF* increased slightly for the focal ability examinees and decreased slightly for high focal ability examinees, but the corresponding change in the proportion of false positive DIF flags was only 0.001.

Table 6. Descriptive Statistics for Distributions of Mean Proportion of Flagged Items, MH D-DIF, and Type-I Error Rate (English Test, No DIF Items)

Condition	Value	B DIF		C DIF		B or C DIF		MH D-DIF		Type-I Error Rate	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Test Length	25	0.030	0.010	0.006	0.005	0.036	0.010	0.031	0.083	0.043	0.011
	50	0.028	0.008	0.005	0.005	0.033	0.008	0.027	0.078	0.040	0.006
	75	0.027	0.007	0.005	0.005	0.032	0.008	0.026	0.076	0.039	0.005
Test Difficulty	Low	0.026	0.006	0.008	0.005	0.034	0.005	0.062	0.111	0.037	0.005
	Middle	0.028	0.008	0.005	0.004	0.033	0.008	0.020	0.057	0.040	0.006
	High	0.031	0.011	0.003	0.003	0.034	0.012	0.002	0.037	0.045	0.010
Test Discrimination	Low	0.027	0.007	0.005	0.004	0.032	0.009	0.022	0.055	0.040	0.006
	Middle	0.028	0.008	0.005	0.005	0.033	0.009	0.028	0.076	0.040	0.007
	High	0.030	0.010	0.006	0.006	0.036	0.009	0.034	0.100	0.042	0.010
Sample Size	100	0.028	0.008	0.009	0.004	0.037	0.006	0.039	0.086	0.037	0.006
	200	0.032	0.008	0.005	0.005	0.037	0.008	0.025	0.078	0.042	0.008
	300	0.024	0.007	0.003	0.004	0.027	0.009	0.020	0.072	0.044	0.008
Focal Ability	Low	0.034	0.009	0.003	0.003	0.037	0.010	-0.041	0.011	0.046	0.010
	Middle	0.025	0.007	0.004	0.003	0.029	0.009	0.011	0.017	0.038	0.004
	High	0.026	0.006	0.009	0.005	0.035	0.006	0.114	0.076	0.039	0.006

Table 7. Mean MH D-DIF by Test Difficulty and Focal Ability (English Test, No DIF Items)

	Focal Ability		
Test Difficulty	Low	Middle	High
Low	-0.044	0.027	0.202
Middle	-0.039	0.007	0.093
High	-0.040	-0.002	0.049

Table 8. Descriptive Statistics for Distributions of Mean Proportion of Flagged Items, MH D-DIF, and Type-I Error Rate (Math Test, No DIF Items)

Condition	Value	B DIF		C DIF		B or C DIF		MH D-DIF		Type-I Error Rate	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Test Length	25	0.037	0.018	0.006	0.004	0.043	0.019	0.020	0.062	0.051	0.021
	50	0.031	0.011	0.005	0.004	0.036	0.011	0.016	0.059	0.043	0.012
	75	0.029	0.009	0.004	0.003	0.034	0.009	0.014	0.056	0.042	0.009
Test Difficulty	Low	0.027	0.007	0.006	0.003	0.033	0.006	0.032	0.078	0.038	0.007
	Middle	0.032	0.012	0.005	0.004	0.037	0.012	0.015	0.053	0.045	0.013
	High	0.038	0.018	0.004	0.004	0.042	0.019	0.003	0.032	0.053	0.021
Test Discrimination	Low	0.028	0.008	0.005	0.003	0.032	0.008	0.020	0.056	0.041	0.008
	Middle	0.031	0.012	0.005	0.004	0.036	0.012	0.017	0.061	0.044	0.014
	High	0.038	0.017	0.006	0.004	0.044	0.018	0.013	0.059	0.051	0.021
Sample Size	100	0.031	0.010	0.008	0.003	0.039	0.010	0.020	0.065	0.039	0.010
	200	0.037	0.014	0.004	0.003	0.041	0.015	0.016	0.058	0.046	0.014
	300	0.029	0.015	0.003	0.002	0.032	0.016	0.013	0.053	0.051	0.019
Focal Ability	Low	0.043	0.016	0.004	0.003	0.047	0.017	-0.041	0.015	0.056	0.021
	Middle	0.025	0.006	0.003	0.003	0.029	0.008	0.004	0.014	0.038	0.004
	High	0.029	0.009	0.008	0.004	0.037	0.009	0.086	0.040	0.042	0.011

Table 9. Mean MH D-DIF by Test Difficulty and Focal Ability (Math Test, No DIF Items)

	Focal Ability		
Test Difficulty	Low	Middle	High
Low	-0.052	0.017	0.132
Middle	-0.042	0.003	0.083
High	-0.028	-0.007	0.043

Table 10. Detection Rate with 10% of DIF items

Condition	Value	Detection Rate (English)	Detection Rate (Math)
Test Length	25	0.676	0.701
	50	0.704	0.713
	75	0.687	0.681
Test Difficulty	Low	0.693	0.729
	Middle	0.718	0.709
	High	0.657	0.658
Test Discrimination	Low	0.560	0.579
	Middle	0.697	0.712
	High	0.811	0.805
Focal Sample Size	100	0.659	0.673
	200	0.735	0.744
	300	0.674	0.679
Focal Ability	Low	0.786	0.813
	Middle	0.639	0.612
	High	0.643	0.671

Discussion and Conclusions

This study identified associations between DIF statistics and other psychometric properties of ACT field test items. There were several notable trends in results. Across content areas, easier items were more likely to exhibit DIF favoring females, whereas harder items were more likely to exhibit DIF favoring males. These findings appear to be novel in the research literature, so they call for replication and further investigation to explain the patterns in DIF results. Only on the science test was it clear that easier items tended to favor White examinees and harder items tended to favor minority examinees. This trend is consistent with prior research in which harder items were more likely to favor Black examinees and easier items were more likely to favor White examinees (e.g., Freedle & Kostin, 1990; Santelices & Wilson, 2012). However, it is difficult to make a direct comparison since the minority focal group in ACT field test analyses includes several racial/ethnic groups and earlier research focused on verbal reasoning items.

There was a slight tendency for higher discrimination items to favor males and lower discrimination items to favor females. The relationship between discrimination and DIF was stronger in the race/ethnicity analysis. Namely, items with higher discrimination were more likely to favor White examinees, and items with lower discrimination were more likely to favor minority examinees. A possible explanation is that, on items with lower discrimination, average items scores (proportion correct) for White and minority

examinees would tend to be more similar. For example, White examinees may be 10% more likely to respond correctly on items with typical (moderate) discrimination. If that difference shrinks to 5% on a small number of items with low discrimination, it would appear that minority examinees perform unusually well on those items (i.e., DIF favoring minority examinees). On highly discriminating items, performance differences could expand (e.g., to 15%), which would appear as DIF favoring White examinees.

A simulation study was conducted to further investigate what assessment conditions might be associated with true positive and false positive DIF classifications. Using item parameters for the ACT English and math tests, a total of 243 simulation conditions were replicated 100 times with varying test length, difficulty, discrimination, focal group size, and focal group ability. The simulation study was repeated with zero DIF items and 10% simulated DIF items. As indicated by results, the proportions of non-DIF items flagged by the MH procedure and *MH CHISQ* Type-I error rates were at or below the expected level of 0.05. Test difficulty, focal group ability, and their interaction had the greatest impacts on DIF results, but those effects were quite small. When simulated DIF was included in each test, the majority of true DIF items were detected, and the effect of simulated DIF on false positive DIF classifications was negligible. Analyses did, however, indicate that detection rates were higher on average for tests with highly discriminating items and when focal group ability was low compared to the reference group.

It might have been useful to know how the characteristics of items and examinees relate to DIF, but this study failed to identify assessment conditions that resulted in false positive DIF classifications beyond the expected rate of 5%. Thus, from a statistical perspective, the MH DIF approach appeared to function appropriately. That may seem like a positive result. However, from the perspective of content developers, that could result in hundreds of field test items each year with false positive DIF classifications. In the empirical analyses reported here, the percentages of field test items flagged for gender and race/ethnicity DIF were each 6%, so most of the significant DIF results could reflect Type-I error.

Eliminating all items flagged for DIF from the item pool would be unacceptable and is not supported by results of this study. Rather, this study supports the current practice of reviewing items flagged for DIF in field test analyses. When an item is flagged for DIF, it seems reasonable to permit the use of that item on future operational tests under the following conditions: the magnitude of the DIF is slight to moderate (e.g., B- or B+ in the MH DIF procedure), other psychometric properties of the item fall within acceptable ranges, and content reviewers cannot identify any reason the item might be biased.

In future DIF analyses, several steps might be considered to reduce the number of false positive DIF classifications. For example, statistical flagging criteria could be modified in a manner that would reduce Type-I error rate (e.g., require $p < 0.01$ for statistical tests or set *MH D-DIF* thresholds higher). Of course, such changes would also reduce the rate at which true DIF is accurately detected. Another possibility to consider is running multiple DIF analysis methods and focusing attention on items flagged by multiple methods.

References

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Fidalgo, Á. M., Mellenbergh, G. J., & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online, 5*(3), 43–53.

Freedle, R., & Kostin, I. (1990). Item difficulty of four verbal item types and an index of differential item functioning for Black and White examinees. *Journal of Educational Measurement, 27*(4), 329–343. https://doi.org/10.1111/j.1745-3984.1990.tb00752.x

Freedle, R., & Kostin, I. (1997). Predicting black and white differential item functioning in verbal analogy performance. *Intelligence, 24*(3), 417–444. https://doi.org/10.1016/S0160-2896(97)90058-1

Freedle, R., Kostin, I., & Schwartz, L. M. (1987). *A comparison of strategies used by Black and White students in solving SAT verbal analogies using a thinking aloud method and a matched percentage-correct design* (RR-87-48). Princeton, NJ: ETS. https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2330-8516.1987.tb00252.x

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

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4), 719–748. https://doi.org/10.1093/jnci/22.4.719

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*(2), 443–451. https://doi.org/10.1177/0013164492052002020

Santelices, M. V., & Wilson, M. (2012). On the relationship between differential item functioning and item difficulty: An issue of methods? Item response theory approach to differential item functioning. *Educational and Psychological Measurement, 72*(1), 5–36. https://doi.org/10.1177/0013164411412943

Scherbaum, C. A., & Goldstein, H. W. (2008). Examining the relationship between race-based differential item functioning and item difficulty. *Educational and Psychological Measurement, 68*(4), 537–553. https://doi.org/10.1177/0013164407310129

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201–210. https://doi.org/10.1177/014662168300700208

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

Qiao Lin

Qiao Lin is an advanced doctoral student in Measurement, Evaluation, Statistics, and Assessment (MESA) at the University of Illinois-Chicago. Qiao's research focuses on using psychometrics to help inform assessment development and implementation as well as applying statistical methods to measure complex learning outcomes.

Jeffrey Steedle, PhD

Jeffrey Steedle is a lead psychometrician in Assessment Transformation directing the team responsible for statistical analyses for the ACT test and guiding research studies related to maintaining measurement quality while making changes to the assessment program. Jeff holds advanced degrees in education, statistics, and educational psychology, and his research interests include assessment validation and motivation on achievement tests.