Research Report

2023-06

What's the DIF?

Item Properties Associated With DIF on the ACT®

JEFFREY T. STEEDLE, SHALINI KAPOOR, AND SHICHAO WANG

Conclusions

In this study, novel statistical approaches helped identify item content characteristics, psychometric properties, and item context variables associated with differential item functioning (DIF), which is a statistical indicator of potential item bias. For example, female students performed differentially well on ACT® English and reading items related to literary narrative texts and when reading test passages included female representation; male students performed differentially well when English and reading passages included scientific content. Black and Hispanic students performed differentially well on earlier items on the English and reading sections; White students performed differentially well on items near the ends of these sections. In addition, White and male students performed differentially well on math word problems, especially those with a real-world context, compared to the focal groups (Asian, Black, Hispanic, and female). On the ACT science section, units involving biology content were differentially easy for female students, whereas units involving physics content were differentially easy for male students. Also, key (i.e., correct response—A, B, C, or D) was sometimes identified as an important predictor of DIF.

So What?

DIF is an indicator of potential item bias, which is why items flagged for DIF are reviewed carefully by diverse panels of content experts. However, DIF can be caused by many other factors such as differences in opportunity to learn or high school course selection. DIF might also be related to differences in student characteristics or behaviors (e.g., motivation, guessing, or omitting), or it may be a statistical artifact caused by DIF methodology. This study helped generate hypotheses to test in future research about the items that contribute most to observed differences in average performance between examinee groups.

Now What?

If DIF is found to be associated with construct-irrelevant factors, item and test development practices might be updated to minimize the risk of DIF. Moreover, results like these presented here might be used to preemptively “neutralize” DIF, for example with reading passages that include female representation (favors female) and cover science topics (favors male). Finally, DIF trends might be examined as guides for addressing systematic differences in achievement within content domains. For example, certain student groups may perform relatively well in some aspects of mathematics because they have better access to instruction in the associated skills.

About the Authors

Jeffrey T. Steedle, PhD

Jeffrey Steedle is a Principal Solutions Designer in the Research department at ACT, Inc., where he conceives of and implements research and development projects for large-scale standardized testing programs. He holds a doctorate in Educational Psychology as well as master’s degrees in Statistics and Education. His research interests include testing motivation, item difficulty modeling, score comparability across testing contexts, and differential item functioning.

Shalini Kapoor, PhD

Shalini Kapoor is a Senior Psychometrician in the Research department at ACT, Inc. Her work involves finding psychometrically sound solutions to maintaining the ACT testing program and supporting new initiatives. Her interests include equating, automated test assembly, mode comparability, process improvement, and computerized adaptive testing.

Shichao Wang, PhD

Shichao Wang is a Psychometrician II in the Research department at ACT, Inc., where she delivers comprehensive psychometric support for large-scale standardized testing programs. She earned both her doctorate and master's degrees in Educational Measurement and Statistics from the University of Iowa. Her research interests include equating and linking, item difficulty modeling, and computer adaptive testing.

This paper was presented during the virtual portion of the 2023 Annual Meeting of the National Council on Measurement, which occurred March 28–30.

Executive Summary

Differential item functioning (DIF) is statistical evidence that a group of examinees performed unusually well or poorly on a given item, and it is commonly interpreted as evidence of potential item bias. With an understanding of the types of items that tend to exhibit DIF, test developers can possibly adjust item-writing and test-construction procedures to minimize sources of construct-irrelevant difficulty for certain examinee groups. With a data set including thousands of items, this study used machine learning to identify important predictors of DIF on the English, math, reading, and science sections of the ACT® test. The available predictors reflected content (subject of a passage; representation of ethnicities, genders, or regions), alignment to content standards, psychometric properties (difficulty and discrimination), and item context (position and key). DIF statistics were calculated to examine differential performance between Asian and White students, Black and White students, Hispanic[1] and White students, and male and female students. Across analyses, the predictors accounted for 0%–40% of the variance in Mantel-Haenszel D-DIF statistics, with the most variance accounted for on the math and English sections.

Several results from this study corroborated prior DIF research. For example, there was often a negative correlation between DIF and difficulty (item proportion correct). That is, easier items were differentially easy for White students, and more difficult items were differentially easy for Black and Hispanic students. Moreover, female students performed differentially well on English and reading items related to literary narrative texts and when reading section passages included female representation; male students performed differentially well when English and reading passages included scientific content. In addition, White and male students performed differentially well on math word problems, especially those with a real-world context, compared to the focal groups (Asian, Black, Hispanic, and female). This study also generated several new results. On the ACT science section, units involving biology content were differentially easy for female students, whereas units involving physics content were differentially easy for male students. Analyses indicated that Black and Hispanic students performed differentially well on earlier items on the English and reading sections; White students performed differentially well on items near the ends of those sections. Also, key (i.e., correct response—A, B, C, or D) was sometimes identified as an important predictor of DIF.

[1] When students register for the ACT, they have the option of providing information about their racial/ethnic background. Students indicate yes, no, or prefer not to respond to the prompt “Indicate if you are of Hispanic or Latino background.” The second prompt states, “Indicate your race. Mark all that apply. (Leave blank if none of these apply to you.)” The options are American Indian/Alaska Native, Asian, Black/African American, Native Hawaiian/other Pacific Islander, White, and prefer not to respond. In this report, we refer to some of these groups as Asian, Black, Hispanic, and White.

This study provided a systematic approach to studying associations between item characteristics and DIF. This information will be used to guide further investigation of construct-irrelevant predictors of DIF such as key. Moreover, these results might inform ways to "neutralize" DIF, such as creating items with "conflicting" features. For example, a reading passage might have science content (favors males) but include female representation (favors females). Note that DIF is an indicator of potential item bias, which is why items flagged for DIF are reviewed carefully, but DIF can be caused by many other factors—such as differences in opportunity to learn or high school course selection and differences in other student characteristics or behaviors (e.g., motivation, guessing, or omitting)—or DIF may be a statistical artifact caused by DIF methodology or measurement models. Further study is planned to investigate the extent to which DIF manifests these factors. For example, if DIF is found to be associated with course-selection patterns, DIF analysis results might serve as a guide to ensuring that all students have equal opportunity to acquire academic knowledge and skills important for college and career readiness.

Background

Though legacy test-development practices in large-scale testing programs are intended to minimize potential item bias, they may actually cause item bias (Randall, 2021). This possibility highlights the need to give more careful consideration to DIF analyses, which are intended to identify potentially biased items. Often, diverse panels of subject matter experts review items flagged for DIF, but these experts hardly ever identify any explanation for the DIF, possibly because item-development and -review processes lead to decontextualized and allegedly culture-free items that eschew item contexts that may be more familiar, engaging, or upsetting to different student groups. This challenge is exacerbated by the many possible causes of DIF. Although DIF statistics are intended as indicators of potential item bias, they can also reflect differential opportunity to learn or course-selection patterns, differences in other student characteristics or behaviors (e.g., motivation, guessing, or omitting), Type-I errors, and statistical artifacts caused by DIF detection methods and measurement models.

Content developers care deeply about test fairness, and they do not want their items to exhibit DIF. At the same time, they do not want to discard good items because of DIF that is unrelated to issues of fairness tied directly to item writing. Subject matter experts may attempt to identify types of items that tend to exhibit DIF, but this is difficult when only a small number of items exhibit statistically significant DIF. As an example, consider an ACT reading passage with 10 associated items. None of those items exhibit statistically significant DIF, yet each item, to a small degree, is differentially easy for White students compared to Black students. Thus, there may be something about that passage related to DIF, and that type of information would be missed when examining only items with statistically significant DIF. Moreover, the small but systematic DIF on those 10 items would have some aggregate relationship to test scores. The goal of this study was to identify content, psychometric, and item-context variables associated with DIF on the English, math, reading, and science sections of the ACT test. With this knowledge, stakeholders can better understand which item types contribute most to observed differences in average scores between gender and racial/ethnic groups. Moreover, if construct-irrelevant factors are found to be associated with DIF, perhaps test developers can manipulate those factors to reduce DIF, thereby reducing construct-irrelevant contributions to achievement differences.

Prior Research

There is a limited body of prior research on item properties associated with DIF, and many of those studies focused narrowly on analogy items. For example, numerous studies identified factors associated with DIF on SAT and GRE analogy items (Freedle et al., 1987; Freedle & Kostin, 1988, 1990, 1991, 1997). Among others, the factors included item difficulty, concreteness, science content, social/personality content, and word frequency. Freedle and his colleagues repeatedly observed a correlation between item difficulty and DIF in Black-White DIF analyses. Specifically, Black examinees performed differentially well on difficult items, and White examinees performed differentially well on easier items. To explain that correlation, the authors proposed a “cultural familiarity” hypothesis. That is, Black and White examinees had different interpretations of “easy” or “more familiar” words used in “everyday conversation,” which led to White examinees performing differentially better on easy items and Black examinees performing differentially better on difficult items.

The DIF-difficulty correlation has been observed on many types of verbal items, and the correlation tends to be lower for items that provide more context (sentence completion and reading comprehension vs. analogy and antonym items; Freedle & Kostin, 1988, 1990). The correlation was also observed on math sections, but it was not as strong (Kulick & Hu, 1989). Some researchers have presented evidence that the DIF-difficulty correlation is related to differential omitting (Kulick & Hu, 1989; Schmitt & Bleistein, 1987). Other researchers demonstrated that the DIF-difficulty correlation persists when matching on item response theory (IRT) ability estimates (rather than raw scores) and when accounting for differences in random guessing behavior (Santelices & Wilson, 2012). More recently, however, Bolt and Liao (2021) demonstrated that the DIF-difficulty correlation is possibly an artifact of negatively asymmetric item characteristic curves, which are expected when disjunctively interacting latent processes underlie item performance. For example, if intended problem-solving processes and proficiency-related guessing both underlie item performance, a correlation between DIF and difficulty would be expected. Correlations between DIF and item discrimination indices have also been observed (Burton & Burton, 1993), though D. M. Bolt (personal communication, April 22, 2022) explained and demonstrated how that correlation could arise from inadequate sum score matching. This issue might be addressed by matching groups in a DIF analysis using IRT ability estimates rather than raw scores.

In other DIF research, Black-White DIF analyses indicated that “verbal” math items favored White examinees and purely numeric math problems favored Black examinees, and the DIF was not an artifact of mean differences between the groups (Rogers & Kulick, 1987; Shepard et al., 1984). Likewise, Asian-White DIF analyses indicated that “verbally-loaded” math items favored White examinees and “pure” math items favored Asian examinees (Kulick & Dorans, 1983), but those effects were smaller when analyzing only data from examinees whose best language was English (Bleistein & Wright, 1987). Carlton and Harris (1992) conducted a thorough study of factors potentially associated with DIF on SAT items. Example findings included sentence-correction items favoring females; emotive, science, and practical-affairs content favoring males; human-relations content favoring female, Black, and Hispanic examinees; reference to minorities favoring Asian, Black, and Hispanic examinees; and science reading passages favoring males. The current study builds on prior research with more recent data, larger samples of items, new content areas (science), and new methods for identifying important predictors of DIF (regression trees).

Method

Measure

The ACT is an educational achievement test designed to measure student mastery of knowledge and skills taught in high school that are recognized as important aspects of college and career readiness. The test is most often used for college admissions and by states to meet federal accountability testing requirements. The test includes four multiple-choice sections (English, math, reading, and science) and an optional writing section. Details about each test section can be found in the *ACT Technical Manual* (ACT, 2022).

Data

For this study, data were analyzed from 46 ACT forms administered during equating studies between October 2020 and February 2022. This included 3,450 English items, 2,760 math items, 1,840 reading items, and 1,840 science items. Mantel-Haenszel DIF analyses (Holland & Thayer, 1986) were conducted with a minimum sample size per group of 150, and a single purification step was included because the matching variable (raw score) can be contaminated by items with DIF. In the purification process, items with significant DIF were removed from the matching variable, and then the DIF analysis was run again with the revised matching variable to produce the final DIF statistics. The following DIF analyses were conducted: Asian-White, Black-White, Hispanic-White, and female-male (each based on self-reported race/ethnicity and gender).

Sample

The sample included 113,799 students, averaging out to approximately 2,400 students taking each test item. The data analyzed for this study came from equating studies, and when ACT forms are equated, the sample is intentionally selected to be representative of ACT examinees nationwide. Based on self-reported demographics, the sample was 6% Asian, 12% Black/African American, 17% Hispanic/Latino, 57% White, 4% two or more races, 4% prefer not to respond, and less than 1% of other racial/ethnic groups (see Footnote 1). In terms of gender, the sample was 42% male, 57% female, less than 1% other, and less than 1% prefer not to respond. The average ACT Composite score (mean of English, math, reading, and science) was 21.0 on the 1–36 scale. For comparison, the national average was 20.3 in 2021.

Analysis

Statistical analyses identified important predictors of the Mantel-Haenszel (MH) D-DIF statistic using regression trees (Breiman, 2001). Specifically, random forests of 1,000 conditional trees were fit to the data sets. This approach has numerous advantages over multiple regression methods of variable selection: (a) having many predictors (even redundant ones) is not a challenge, (b) there are no assumptions about normality or linearity, (c) it identifies the best predictors (and interactions among them) automatically, (d) it produces importance statistics for predictor variables, and (e) cross-validation is built in. The available predictors reflected content (subject of a passage; representation of ethnicities, genders, or regions), alignment to content standards, psychometric properties (p-value and point-biserial correlation), and item context (position and key). A complete list and description of predictors are provided in Appendix A. The regression tree analyses identified the most important predictors and estimated the proportion of MH D-DIF variance accounted for by the predictors (R^2).

Descriptive analyses were then conducted on important predictors. This involved calculating the mean MH D-DIF value for groups of items to evaluate the extent to which those items were differentially easy or difficult. Note that results were not considered when a certain item type had too few items to support generalization (fewer than 5 items or 1 passage for passage-level predictors of DIF). Recall that DIF-difficulty and DIF-discrimination correlations may be statistical artifacts unrelated to possible item bias. Accordingly, since some predictors were confounded with item difficulty and/or discrimination (e.g., items in a certain reporting category were more difficult on average), mean MH D-DIF residuals after controlling for difficulty and discrimination were also calculated. The regression analyses included second-degree polynomials to allow for nonlinear associations. When DIF was correlated with difficulty and/or discrimination, controlling for difficulty and discrimination tended to shift the DIF statistics, though patterns in results were generally the same before and after controlling; departures from this pattern are noted. Content experts reviewed results to help describe the types of items that tended to exhibit DIF. A summary of their comments follows the statistical results.

Results

DIF Descriptive Statistics

Table 1 provides descriptive statistics for the MH D-DIF distributions. Throughout this report, negative DIF indicates differential difficulty for the focal group (Asian, Black, Hispanic, or female), and positive DIF indicates differential difficulty for the reference group (White or male). For example, negative DIF in the Black-White DIF analysis occurred when items were relatively difficult for Black students compared to White students after matching students in the two groups on overall achievement level. When a type of item is said to have “favored” a certain group, that is another way of saying those items were differentially easier for that group (and differentially difficult for the comparison group).

In the Asian-White DIF analyses, there was an overall tendency for items to be differentially more difficult for White students, which was indicated by the positive mean MH D-DIF statistics. However, the opposite was true in the Black-White and Hispanic-White DIF analyses. That is, there was a tendency for items to be differentially more difficult for Black and Hispanic students compared to White students—slightly more so on the math and science sections compared to the English and reading sections. Across all English and reading items, there was no systematic pattern of items being differentially easy or difficult when comparing male and female students. However, the mean MH D-DIF statistics for math and science items were both negative, indicating a tendency to be differentially more difficult for female students. Note that these mean differences are quite small on the scale of MH D-DIF statistics, where items with MH D-DIF between −1.0 and 1.0 would be considered as exhibiting “negligible or nonsignificant DIF” (Zwick, 2012, p. 2). Analyses reported in subsequent sections indicate whether there were systematic differences between types of items within test sections.

Table 1. Means and Standard Deviations (SD) of MH D-DIF Distributions

Section	Asian-White		Black-White		Hispanic-White		Female-Male	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
English	0.09	0.75	−0.11	0.67	−0.05	0.52	−0.01	0.49
Math	0.07	0.73	−0.15	0.66	−0.07	0.50	−0.04	0.54
Reading	0.05	0.59	−0.09	0.58	−0.04	0.46	0.00	0.46
Science	0.05	0.58	−0.12	0.60	−0.05	0.46	−0.03	0.46

DIF-Difficulty Correlations

As in many prior studies, there was often a significant correlation between MH D-DIF and item difficulty (i.e., proportion correct or “p-value”). In Table 2, a negative correlation indicates that more difficult items (with lower p-values) tended to have more positive MH D-DIF (i.e., favor the focal group), and easier items (with higher p-values) tended to have more negative MH D-DIF (i.e., favor the reference group). In the Asian-White analysis, easier English items tended to favor White students, and harder English items tended to favor non-White students. The strongest DIF-difficulty correlations were observed in the Black-White analysis: −.327 (p < .001) for math and −.345 (p < .001) for science. Across test sections, there was a consistent tendency for harder items to favor Black students and easier items to favor White students. The same was true in the Hispanic-White analyses, though the correlation for reading was not statistically significant. In the female-male analyses, easier items tended to favor female students, and more difficult items tended to favor male students, particularly on the reading and science sections. In several cases described in subsequent sections, the relationship between item difficulty and MH D-DIF was U-shaped (i.e., easier items were generally more likely to exhibit DIF).

Table 2. Correlations Between p-values and MH D-DIF

Section	Asian-White	Black-White	Hispanic-White	Female-Male
English	−.102***	−.157***	−.074***	.043*
Math	−.023	−.327***	−.160***	.050**
Reading	.004	−.058*	−.022	.196***
Science	.029	−.345***	−.216***	.158***

*p < .05. **p < .01. ***p < .001.

R^2 and Variable Importance

Table 3 shows R^2 values based on the random forest analyses. These values reflect cross-validation because they are based on a random set of items withheld when fitting each of the 1,000 conditional trees. The English R^2 values ranged from .12 (Hispanic-White) to .22 (Black-White). The math R^2 values were the highest on average, with a range of .22 (Hispanic-White) to .40 (female-male). For reading items, the available predictors accounted for little variance in Hispanic-White DIF (.08) and practically none of the Asian-White DIF (.01). However, the predictors accounted for notable proportions of the Black-White (.16) and female-male (.22) DIF. The available predictors accounted for some science item DIF in the Black-White (.22), Hispanic-White (.12), and female-male (.15) analyses, but none in the Asian-White analysis.

Even when R^2 was relatively low (e.g., around .10), there were predictors (or certain values of predictors; e.g., alignment to certain content standards or reporting categories) that were associated with DIF. Tables 4–7 provide the importance statistics for the predictors of MH D-DIF. Note that importance statistics are scaled to a maximum of 100 regardless of R^2. For example, strand/topic/standard (i.e., Common Core State Standard alignment) was among the most important predictors in each of the four math analyses (Table 5). Note that, when R^2 is very low (Table 3), there are no important predictors, so the importance statistics should not be interpreted. The most important predictors in each analysis will be considered in the following sections.

Table 3. Conditional Random Forest R^2 Values

Section	DIF analysis	R^2
English	Asian-White	.14
	Black-White	.22
	Hispanic-White	.12
	Female-Male	.17
Math	Asian-White	.24
	Black-White	.34
	Hispanic-White	.22
	Female-Male	.40
Reading	Asian-White	.01
	Black-White	.16
	Hispanic-White	.08
	Female-Male	.22
Science	Asian-White	.00
	Black-White	.22
	Hispanic-White	.12
	Female-Male	.15

Table 4. Variable Importance for Predicting DIF on the English Section

Predictor	Asian-White	Black-White	Hispanic-White	Female-Male
Difficulty (p-value)	11	86	39	16
Discrimination (point-biserial correlation)	4	45	24	4
Position	1	40	48	3
Key	19	28	23	61
Reporting category	28	37	25	48
Depth of knowledge	15	14	17	30
Content standard	100	100	100	100
Passage name	—	47	59	24
Passage position	—	31	25	—
Passage type	—	3	—	3
Passage subtype	—	11	5	2
Gender representation	—	1	—	1
Ethnicity representation	—	4	2	1
Region representation	—	3	3	—
Urbanicity representation	—	2	—	—
Item character count	16	14	9	10
Passage character count	—	—	1	—

Table 5. Variable Importance for Predicting DIF on the Math Section

Predictor	Asian-White	Black-White	Hispanic-White	Female-Male
Difficulty (p-value)	11	100	61	11
Discrimination (point-biserial correlation)	14	61	37	11
Position	11	64	44	15
Key	7	36	41	8
Reporting category	59	41	71	56
Depth of knowledge	11	5	8	17
Strand/topic/standard	82	78	100	100
Advanced standard	10	9	5	4
Real-world context	51	75	57	66
Modeling	10	5	11	17
Item character count	100	35	75	51

Table 6. Variable Importance for Predicting DIF on the Reading Section

Predictor	Asian-White	Black-White	Hispanic-White	Female-Male
Difficulty (p-value)	—	43	29	28
Discrimination (point-biserial correlation)	—	100	100	1
Position	—	48	59	4
Key	—	22	28	3
Reporting category	—	21	14	8
Understanding complex texts	—	4	7	3
Depth of knowledge	—	5	2	3
Content standard	—	32	34	31
Passage name	—	70	78	62
Passage position	—	49	64	6
Passage type	—	3	5	100
Passage subtype	—	14	—	70
Gender representation	—	2	—	37
Ethnicity representation	—	6	6	11
Region representation	—	3	1	11
Urbanicity representation	—	3	—	17
Item character count	—	19	35	5
Passage character count	—	2	7	3

Table 7. Variable Importance for Predicting DIF on the Science Section

Predictor	Asian-White	Black-White	Hispanic-White	Female-Male
Difficulty (p-value)	—	100	83	47
Discrimination (point-biserial correlation)	—	90	100	5
Position	—	7	7	3
Key	—	10	12	3
Reporting category	—	3	—	22
Depth of knowledge	—	2	2	3
Skill standard	—	15	21	100
Passage skill standard	—	2	4	35
Passage name	—	40	60	87
Passage position	—	7	6	2
Passage content	—	1	1	26
Format	—	2	—	5
Background knowledge	—	2	9	3
Background knowledge type	—	3	8	16
Item character count	—	6	7	37
Passage character count	—	—	—	12

English DIF Trends

Asian-White DIF Analysis

Note that the ACT English section measures students' writing and revising skills with multiple-choice items. The available English predictors accounted for 14% of the variance in MH D-DIF in the Asian-White DIF analysis. Content standard alignment was the most important predictor of MH D-DIF. The standards with items that favored White and Asian students the most on average are shown in Table 8. Using idiomatic language (e.g., selecting the appropriate preposition such as *on*, *in*, *at*, *to*, *along*, *near*, etc.) favored White students the most. Items dealing with possessive pronouns favored Asian students the most. Generally, items dealing with punctuation conventions and sentence structure and formation tended to favor Asian students; items dealing with topic development in terms of purpose and focus tended to favor White students.

Table 8. English Content Standards That Favored White and Asian Students

Favored White	Favored Asian
Using idiomatic language (most)	Correcting run-on sentences, including comma splices
Ensuring pronoun-antecedent agreement	Ordering sentences in a logical sequence
Making decisions about paragraph division based on a specified criterion	Correcting rhetorically ineffective sentence fragments
Identifying purpose of specified word, phrase, or sentence	Correcting squinting and dangling modifiers
Ensuring precision of language	Eliminating unnecessary punctuation
Determining whether a text has met a specified primary purpose	Forming possessive nouns
—	Ensuring concision of language
—	Using within-sentence punctuation to indicate sharp breaks
—	Using appropriate pronoun case
—	Forming possessive pronouns (most)

Black-White DIF Analysis

The available English predictors accounted for 22% of the variance in MH D-DIF in the Black-White DIF analysis. Again, content standard alignment was the most important predictor of MH D-DIF. The item classifications that were the most differentially difficult for White and Black students are shown in Table 9. Eliminating unnecessary punctuation items tended to favor Black students the most. As in the Asian-White DIF analysis, using idiomatic expressions favored White students the most. Overall, items dealing with topic development favored White students. Organization, unity, and cohesion items and style items also favored White students, but those relationships were significantly weakened after controlling for difficulty and discrimination. As for reporting categories, the production of writing category favored White students on average, but that relationship was also weakened after controlling for difficulty and discrimination.

Table 9. English Content Standards That Favored White and Black Students

Favored White	Favored Black
Using idiomatic language (most)	Correcting rhetorically ineffective sentence fragments
Ensuring precision of language	Ensuring concision of language
Determining the most logical placement for a sentence in paragraph or text	Correcting squinting and dangling modifiers
Correcting vague or ambiguous pronouns	Determining effects of stylistic choices/using style for effect
Correcting misplaced modifiers	Eliminating unnecessary punctuation (most)
Determining whether a text has met a specified primary purpose	—
Using a word, phrase, or sentence to accomplish a specified purpose	—
Identifying purpose of specified word, phrase, or sentence	—
Maintaining consistency in style and tone	—
Using an effective introductory, concluding, or transition sentence	—
Determining relevance of material in terms of focus	—
Distinguishing between and among frequently confused words	—

Passage name was also an important predictor of MH D-DIF. That is, the items associated with certain passages were differentially difficult on average for Black or White students. This finding could be related to factors such as topic, genre, and skills measured by the items. Note that the racial/ethnic representation variable was not an important predictor of DIF, nor were the other representation variables (gender, region, and urban vs. rural). Overall, passage subtype was not an important predictor of DIF, but passage subtype gave some indication of the types of passages favoring Black and White students. Passages about language favored Black students; passages about chemistry, friends, sociology, and literary criticism favored White students. Key was weakly associated with MH D-DIF. Items with a key of B tended to favor White students. The English language arts (ELA) content developers noted that response option B is often the longest option because the response options are typically ordered from longest to shortest starting at B (option A most often is NO CHANGE).

As reported in Table 2, MH D-DIF correlated −.157 with difficulty in the Black-White analysis. MH D-DIF also correlated −.188 ($p < .001$) with item discrimination (point-biserial correlation) and −.182 with item position ($p < .001$). Note that ACT test sections are constructed to progress from easier to more difficult. For a passage-based section such as English, average item difficulty for passages is used to order the passages. Yet the easier items and items later in the section tended to favor White students. More difficult and earlier items tended to favor Black students.

Hispanic-White DIF Analysis

In the Hispanic-White DIF analysis, the available English predictors accounted for only 12% of the variance in MH D-DIF, so few predictors are described here. Content standard alignment was the most important predictor of MH D-DIF. The standards that favored White and Hispanic students the most are shown in Table 10. Again, using idiomatic language favored White students the most on average. Forming possessive pronouns favored Hispanic students. Overall, topic development in terms of purpose and focus items weakly favored White students. No group of standards systematically favored Hispanic students. Passage name was also an important predictor of DIF, indicating that some passages had items that were differentially difficult for White or Hispanic students on average. Looking closer at passage subtype as an indicator of passage topic, botany and astronomy passages tended to favor White students more than other subtypes.

Table 10. English Content Standards That Favored White and Hispanic Students

Favored White	Favored Hispanic
Using idiomatic language (most)	Determining effects of stylistic choices/using style for effect
Correcting vague or ambiguous pronouns	Forming possessive pronouns (most)
Ensuring precision of language	—
Maintaining consistency in style and tone	—
Determining the most logical placement for a sentence in paragraph or text	—
Using a word, phrase, or sentence to accomplish a specified purpose	—
Identifying purpose of specified word, phrase, or sentence	—

Among the quantitative predictors of MH D-DIF, item position and item difficulty were the most important. Item position correlated −.190 (p < .001) with MH D-DIF. Thus, items earlier in the section tended to favor Hispanic students, and items later in the section tended to favor White students. The association between difficulty and DIF was systematic but U-shaped (i.e., easier items were more likely to exhibit DIF in general), which explains why difficulty was an important predictor, yet it had a low correlation with DIF ($r = -.074$, $p < .001$).

Female-Male DIF Analysis

The available English predictors accounted for 17% of the variance in MH D-DIF in the female-male DIF analysis. Content standard alignment was the most important predictor of MH D-DIF. Table 11 shows the standards that favored female and male students the most on average. The standard that favored females the most dealt with using within-sentence punctuation to indicate sharp breaks. The standard that favored males the most related to ensuring precision of language by selecting the most precise or logical word to complete a sentence. On average, there were weak tendencies for items about sentence structure and formation to favor female students and for items about expressing ideas clearly to favor male students.

Table 11. English Content Standards That Favored Male and Female Students

Favored male	Favored female
Ensuring precision of language (most)	Correcting faulty subordination, coordination, and parallelism
Using idiomatic language	Using punctuation to separate items in a series
Using a word, phrase, or sentence to accomplish a specified purpose	Forming possessive pronouns
Distinguishing between and among frequently confused words	Correcting rhetorically ineffective sentence fragments
Forming possessive nouns	Using within-sentence punctuation to indicate sharp breaks (most)

Key was also associated with female-male MH D-DIF. Items with a key of D (the last and often shortest response option) tended to favor female students; items with a key of A (most often NO CHANGE) tended to favor male students. Overall, passage name was not a strong predictor of female-male MH D-DIF, but there were numerous passages with items that favored males or females on average. A cursory reading of passage titles revealed that many of the passages that favored females the most had an artistic topic, and many of the passages that favored males the most had a scientific topic. An examination of passage subtypes indicated that dance and other personal narrative passages favored female students, while architecture passages favored male students. The passage gender representation variable was not an important predictor of female-male DIF on the English section, but the content experts recognized an association between gender representation coding and DIF at the passage level (see the following Content Expert Observations section).

Content Expert Observations

The ELA content development team focused their review on systematic DIF at the passage level (i.e., passages with items that tended to exhibit differential difficulty for a certain group). Note that ACT English passages are written by the ELA content development team, whereas ACT reading passages are used with permission of the authors. The ELA team observed that the English passages that exhibited the most DIF covered a range of topics, with the standard passage types (humanities, social science, and natural science) appearing in similar numbers. A few possible trends emerged, though the ELA team expressed the desire to get more data and closely examine the characteristics of passages that were not flagged for systematic DIF.

The team recognized that humanities passages with diverse representation, which were mainly arts-focused passages about diverse artists or topics, tended to favor focal groups (Black, Hispanic, and female). The team reported that the humanities passages that they write exhibit greater cultural diversity than the other passage types. Social science passages—especially those with a history subtype—tended to favor the reference groups (White and male). Nearly all natural science passages with systematic DIF favored the reference groups.

Among passages flagged for systematic DIF, passages coded for gender representation often exhibited DIF in the direction of the represented gender. For example, of the 10 passages with systematic DIF and a male gender representation code, nine passages favored male students. Likewise, it was often true that passages with an ethnicity representation code tended to favor a non-White group. For example, of the 11 passages with systematic DIF and an ethnicity representation code in the Hispanic-White DIF analysis, nine passages favored Hispanic students. Four of those passages were coded as Mexican American/Chicano/Latino representation, and three of those passages favored Hispanic students. The ELA team also identified a few passages where the direction of the average DIF was unexpected. For example, items associated with a passage about a book club centered on Black women writers favored White students over Black students, and a passage about a Mexican American woman who collected plant specimens in the Amazon consistently favored the reference groups. Perhaps the passage types—social science and natural science, respectively—had a greater influence on student performance.

Personal narrative passages tended to favor White students over non-White students and female students over male students. There were few personal narrative passages, so the ELA content team expressed a desire for more data. They also expressed a desire to see more data from passages with an international (non-U.S.-based) race/ethnicity code. There were only two such passages in the study's data set.

Math DIF Trends

Asian-White DIF Analysis

In the Asian-White analysis, item character count (including all text, spaces, and mathematical notation) was the most important predictor of MH D-DIF. Character count and MH D-DIF correlated −.365 ($p < .001$), indicating that items with fewer characters tended to favor Asian students, and items with more characters tended to favor White students. Other important predictors included Common Core State Standards (CCSS) strand, reporting category (the parent of strand), and the presence of a real-world context. Items from the reporting categories number and quantity, algebra, and functions tended to favor Asian students. Many of the content strands favoring White students were related to the reporting categories integrating essential skills[2] and statistics and probability, which are more likely to include a real-world

[2] According to the *ACT Technical Manual* (2022), "This reporting category focuses on whether students can put together knowledge and skills to solve problems of moderate to high complexity. Topics include

context and have higher character counts. Those strands are listed in Table 12. Thus, the important MH D-DIF predictors told a consistent story of White students performing differentially well on longer word problems and Asian students performing differentially well on evaluating functions and on shorter items requiring symbolic manipulation.

Table 12. Math Strands That Favored White and Asian Students

Favored White	Favored Asian
Ratios and proportional relationships (most)	Expressing geometric properties with equations
Operations and algebraic thinking	Reasoning with equations and inequalities
Making inferences and justifying conclusions	Trigonometric functions
The number system	Seeing structure in expressions
Congruence	Solving real-world and mathematical problems involving area, volume and surface area of two- and three-dimensional objects composed of triangles, quadrilaterals, polygons, cubes, and right prisms
Number and operations in base 10	Similarity, right triangles, and trigonometry
Conditional probability and the rules of probability	Vector and matrix quantities
Circles	Interpreting functions
Measurement and data	The real number system
Number and operations—fractions	Building functions
Quantities	The complex number system
Using probability to make decisions	Arithmetic with polynomials & rational expressions (most)

Black-White DIF Analysis

The available math predictors accounted for 34% of the variance in MH D-DIF in the Black-White DIF analysis. Item proportion correct (p-value) was the most important predictor of MH D-DIF. With a correlation of −.327 (p < .001), more-difficult items tended to favor Black students, and easier items tended to favor White students. Several other quantitative predictors were also important: item position correlated .304 with MH D-DIF (p < .001), point-biserial correlation correlated −.262 with MH D-DIF (p < .001), and character count correlated −.232 with MH D-DIF (p < .001). Thus, items exhibiting DIF favoring White students tended to have the following

rate and percentage; proportional reasoning; area, surface area, and volume; quantities and units; expressing numbers in different ways; using expressions to represent quantities and equations to capture relationships; rational exponents; the basics of functions; function notation; sequences as functions; transformations, congruence, symmetry, and rigid motions; data analysis and representation; measures of center and spread; normal distribution; associations between two variables; two-way tables; scatterplots; linear models; correlation; and model fit" (pp. 31–32).

properties: easier, occurred earlier in the section (a strong indicator of easiness), had higher discrimination, and had more characters. Items that tended to favor Black students were more difficult, occurred later in the section (a strong indicator of difficulty), had lower discrimination, and had fewer characters.

Content alignment variables strand and reporting category were also important predictors of MH D-DIF. Items in the integrating essential skills and statistics and probability reporting categories and the strands listed in Table 13, which were also those more likely to have a real-world context, tended to exhibit DIF favoring White students. Items in those reporting categories tended to be easier and have higher character counts, so this finding was consistent with the correlations among predictors. Finally, items with a key (correct response) of E favored White students on average. Items with a key of E tend to be more difficult (and therefore appear near the end of the section), but even after controlling for difficulty, there was still a tendency for items with a key of E to favor White students. Note that math is the only section with five response options (English, reading, and science have only four).

Table 13. Math Strands That Favored White and Black Students

Favored White	Favored Black
Ratios and proportional relationships (most)	Vector and matrix quantities
Statistics & probability	Building functions
Quantities	The complex number system
Number and operations—fractions	Expressing geometric properties with equations
Number and operations in base 10	Circles
Making inferences and justifying conclusions	Trigonometric functions
Measurement and data	Arithmetic with polynomials & rational expressions (most)

Hispanic-White DIF Analysis

The available math predictors accounted for 22% of the variance in MH D-DIF in the Hispanic-White DIF analysis. As in the Black-White DIF analysis, items with strands and reporting categories related to integrating essential skills and statistics and probability (Table 14), items with real-world contexts, and items with higher character counts tended to favor White students. Items without a real-world context did not show systematic DIF favoring either group. Algebra items had a weak tendency to favor Hispanic students. Again, items with a key of E tended to favor White students the most, but the average MH D-DIF for such items was reduced substantially after controlling for difficulty and discrimination (items with a key of E were the most difficult on average). MH D-DIF correlated −.230 with character count (p < .001), −.160 with item difficulty (p < .001), .141 with item position (p < .001), and −.122 with item discrimination (p < .001). Thus, in terms of quantitative predictors, items that tended to favor Hispanic students had fewer characters, were more difficult, occurred later in the section (a strong indicator of difficulty), and had lower discrimination. The converse was true for items that favored White students on average.

Table 14. Math Strands That Favored White and Hispanic Students

Favored White	Favored Hispanic
Statistics & probability (most)	The complex number system
Making inferences and justifying conclusions	Seeing structure in expressions
Quantities	Trigonometric functions
Number and operations—fractions	Arithmetic with polynomials & rational expressions (most)
The number system	—
Number and operations in base 10	—
Measurement and data	—
Operations and algebraic thinking	—
Ratios and proportional relationships	—

Female-Male DIF Analysis

The available math predictors accounted for more variance in the female-male DIF analysis than any other analysis (40%). Much of that variance was accounted for because items related to the reporting categories statistics and probability and integrating essential skills, which are also items more likely to have real-world contexts and have higher character counts, tended to favor males. Items in the reporting categories algebra and number and quantity tended to favor female students. The math strands that favored male and female students the most are listed in Table 15. MH D-DIF correlated −.359 with character count (p < .001). These trends did not appear to be explained by any other factors such as item difficulty.

Table 15. Math Strands That Favored Male and Female Students

Favored male	Favored female
Ratios and proportional relationships (most)	Vector and matrix quantities
Making inferences and justifying conclusions	Expressions and equations
Quantities	Interpreting functions
Number and operations in base 10	Building functions
Measurement and data	Reasoning with equations and inequalities
The number system	Trigonometric functions
Statistics & probability	The real number system
Ratios and proportional relationships	Arithmetic with polynomials & rational expressions
Operations and algebraic thinking	Seeing structure in expressions
Conditional probability and the rules of probability	The complex number system (most)
Using probability to make decisions	—

Content Expert Observations

The math content expert team mainly focused on the relationships among the important predictors of DIF. For example, p-value and item position were very strongly related because ACT math sections are constructed to progress from easier to more difficult items. Having a real-world context, content standard alignment, character count, and difficulty are all expected to relate. For example, items in the integrating essential skills reporting category are more likely to have real-world contexts, have higher character counts, and be easier on average.

The math team expressed interest in studying the interaction between character count and context to disentangle whether the context or the item length was the main issue. Supplemental regression analyses applied to the Black-White and female-male data sets helped inform this issue. In results, character count was substantially weakened as a predictor of MH D-DIF when controlling for the presence of a real-world context. Thus, the presence of a real-world context (and the types of skills measured in those contexts) appeared to be a more important predictor of DIF than item length. Having a real-world context had a negative regression coefficient (i.e., such items tended to favor the reference groups), but that effect was smaller for geometry and statistics and probability items.

The math team also noted that items toward the ends of math sections tend to have more D and E keys, which might explain why the key of E was sometimes related to DIF favoring White students. That explanation was inconsistent with harder items tending to favor Black students, but the result could also have reflected differential guessing or omitting at the very end of the section. The team also theorized that this finding might have something to do with nonnumerical response options sometimes being ordered from shortest to longest. Another theory was that students find an attractive (incorrect) response among options A–D when the key is E. The team noted that computational items, particularly those with numeric answers, are more likely to have keys of B, C, or D because item developers typically write distractors with numeric values higher and lower than the key (and item responses usually are in numerical order). More-conceptual items that require consideration of all the response options (and choosing the best option) are therefore more likely to have a key of E.

Reading DIF Trends

Asian-White DIF Analysis

The available reading predictor variables accounted for only 1% of the variance in MH D-DIF in the Asian-White DIF analysis. That is, there were no important predictors of MH D-DIF in this analysis.

Black-White DIF Analysis

The available reading predictors accounted for 16% of the variance in MH D-DIF in the Black-White DIF analysis. Item discrimination (point-biserial correlation) was the most important predictor of MH D-DIF ($r = -0.266$, $p < .001$). Item position correlated −.179 with MH D-DIF

(p < .001), and passage position was also an important predictor. Thus, items that were less discriminating and earlier in the section tended to favor Black students, and items that were more discriminating and later in the section tended to favor White students. Item difficulty (p-value) had a U-shaped association with MH D-DIF (i.e., easier items were more likely to exhibit DIF).

After item discrimination, passage name was the second most important predictor of MH D-DIF, indicating that items associated with certain passages tended to favor Black or White students. Rather than attempting to conduct a qualitative analysis of passage titles and content, passage subtype was analyzed to obtain some indication of the passage topics that favored Black or White students. Note that passage subtype was not an important predictor overall, but several subtypes had systematic tendencies to exhibit DIF. The subtypes that favored Black students the most were anthropology, short story, and environmentalism. The subtypes that favored White students the most were business, astronomy, biology, and zoology.

Content standard alignment was a moderately important predictor of MH D-DIF. The standards that tended to favor White and Black students the most are shown in Table 16. Items that involved locating important details in the passage favored Black students the most on average, and items that required determining the meaning of words and phrases from context tended to favor White students the most. Overall, word meanings and word choice and also visual and quantitative information items tended to favor White students.

Table 16. Reading Content Standards That Favored White and Black Students

Favored White	Favored Black
Determining the meaning of words & phrases from context (most)	Locating important details
Determining the meaning of figurative language	—
Determining stated & implied main ideas/themes of whole texts	—

Hispanic-White DIF Analysis

The available reading predictors accounted for only 8% of the variance in MH D-DIF in the Hispanic-White DIF analysis. As in the Black-White DIF analysis, item discrimination (point-biserial correlation) was the most important predictor of MH D-DIF (r = –.202, p < .001). Item position correlated significantly with MH D-DIF (r = –.179, p < .001), and passage position (1–4) was also an important predictor. Thus, items with lower levels of discrimination or placed earlier in the section favored Hispanic students more on average. Conversely, items that exhibited higher levels of discrimination or were positioned toward the end of the section were differentially easy for White students.

Passage name was an important predictor of MH D-DIF. Based on passage subtype, psychology passages tended to favor Hispanic students. Ecology, zoology, and natural history passages tended to favor White students. Finally, item character count was a moderately important predictor of MH D-DIF. The correlation with MH D-DIF was relatively low (r = .068, p < .01), but there was a clear tendency for the shortest items to favor White students. The shortest items tended to be vocabulary items, and items measuring vocabulary standards also appeared first in the list of standards that favored White students the most on average (Table 17). Items requiring the synthesis of multiple texts relating to central ideas, themes, and summaries tended to favor Hispanic students. Overall, word meanings and word choice items favored White students the most. No major content grouping of items tended to favor Hispanic students.

Table 17. Reading Content Standards That Favored White and Hispanic Students

Favored White	Favored Hispanic
Determining the meaning of words & phrases from context (most)	Synthesis of multiple texts in paired units (most)
Determining the meaning of figurative language	—
Distinguishing among fact, opinion, reasoned judgment, & value judgment	—
Analyzing the function of specific words	—

Female-Male DIF Analysis

The available reading predictors accounted for more variance in the female-male DIF analysis than in any other reading analysis (22%). Passage type was the most important predictor of MH D-DIF in the female-male DIF analysis. On average, natural science and social science passages favored male students, whereas literary narrative passages favored female students. Passage subtype, which is nested within passage type, and passage name were the next most important predictors of MH D-DIF. Even after controlling for difficulty and discrimination, female students performed differentially well on personal essay, novel, short story, and memoir/autobiography passages as well as on passages relating to theater, film, and literary criticism. Male students performed differentially well on passages relating to physics, ecology, astronomy, psychology, technology, architecture, and natural science. The gender representation indicator was an important predictor. Passages with female representation tended to favor female students, but passages with male representation did not systematically favor either gender group.

Reading content standard alignment was a moderately important predictor of MH D-DIF. On average, items that required understanding the point of view in narrative texts favored females most, and items that required determining the meaning of words and phrases from context favored males most (Table 18). Generally, visual and quantitative information, word meanings and word choice, and arguments items were more likely to favor male students. No group of standards systematically favored female students. Item difficulty was also associated with MH D-DIF (r = .196, p < .001). Specifically, easier items tended to favor female students, and more difficult items had a weak tendency to favor male students. Note that female students, on average, perform better on the ACT reading section than male students.

Table 18. Reading Content Standards That Favored Male and Female Students

Favored male	Favored female
Determining the meaning of words & phrases from context (most)	Synthesis of multiple texts relating to textural evidence
Distinguishing among fact, opinion, reasoned judgment, & value judgment	Determining stated & implied main ideas/themes of whole texts
Identifying & analyzing textual evidence to support claims & counterclaims	Understanding point of view in narrative texts (most)

Content Expert Observations

The ELA content expert team mainly focused on examining systematic DIF among reading passage types. Additionally, the team looked for associations between DIF and ethnicity coding, which is assigned during test development when a passage focuses on individuals who identify as a certain ethnicity.

It was noted that passages with items exhibiting systematic DIF covered a wide range of topics and approaches, with the standard types (literary narrative, social science, humanities, natural science, and paired passages) appearing in similar numbers. Note that paired passages could be any standard type of passages, and both passages would be in the same category. A few possible trends were discernible—including associations between ethnic or gender representation and DIF favoring the represented group—though more data should be collected and reviewed before drawing any firm conclusions.

Although the ethnic representation variable was not an important predictor of DIF in the statistical analyses, the ELA team observed possible trends when examining certain passage types. In general, items associated with literary narrative passages favored Black, Hispanic, and female students. Six out of the 10 literary narrative passages with items systematically favoring Black students were coded for ethnic representation. Likewise, seven out of 10 such passages with items that systematically favored Hispanic students were coded for ethnic representation. Female students were overwhelmingly favored on items associated with literary narrative passages. Out of 27 literary narrative passages that exhibited systematic DIF, 26 of them were differentially easy for female students. Fourteen of those 26 passages were coded by female representation, and 13 were coded for ethnic representation.

Items connected to social science passages were differentially easy for White students compared to Black and Hispanic students and also for male students compared to female students. There were no notable trends in social science passages coded for ethnic representation, but 10 out of the 17 passages with items that systematically favored males were coded for male representation.

Items associated with humanities passages favored White students compared to Black and Hispanic students. These items also tended to favor female students compared to male students. Among 12 passages with items that systematically favored White students compared to Black students, five were coded for ethnic representation. All six paired humanities passages had items that were differentially easy on average for White students compared to Black students. Additionally, out of the eight passages with items that systematically favored White students compared to Hispanic students, three were coded for ethnic representation.

Items associated with natural science passages were differentially easy for White students compared to Black and Hispanic students. Overwhelmingly, males performed differentially well on natural science passages compared to females. Indeed, all 20 natural science passages with items that exhibited systematic DIF favored males.

In general, items related to paired passages tended to favor White students compared to Black students, yet these items favored Hispanic students compared to White students. These items also favored female students compared to male students. A small number of social science, humanities, and natural science passages also included visual and quantitative information items (e.g., requiring the interpretation of tables or figures). Such passages tended to favor the reference groups (White and male), but very few of these were represented in the data.

Science DIF Trends

Asian-White DIF Analysis

The available science predictor variables accounted for none of the variance in MH D-DIF in the Asian-White DIF analysis. That is, there were no important predictors of MH D-DIF in this analysis.

Black-White DIF Analysis

The available science predictors accounted for 22% of the variance in MH D-DIF in the Black-White DIF analysis. Item proportion correct (p-value) was the most important predictor of MH D-DIF ($r = -.345$, $p < .001$). It was followed closely by item discrimination ($r = -.367$, $p < .001$). That is, more difficult and less discriminating items tended to favor Black students, and easier and more discriminating items tended to favor White students.

Passage name was also a moderately important predictor of MH D-DIF. However, the science content expert team was unable to discern any patterns in the relationship between passage name and DIF. They hypothesized that correlations between passage name and DIF could be due to other factors such as passage content standard. Upon closer inspection of other predictors, passages with physics content tended to favor White students, especially items related to passage-level standards for dynamics and electric and magnetic fields. These two passage standards accounted for only 8.6% of all science items, which could explain why passage standard was not identified as an important predictor.

There was also a tendency for items in the scientific investigation (SIN) reporting category to favor White students, though the relationship was notably weakened after controlling for difficulty and discrimination. For example, White students performed differentially well on items related to the following skills: evaluate the design or methods of an experiment, perform an extrapolation using data in a table or graph, find information in the text that describes a data presentation, determine the experimental conditions that would produce specified results, and predict the effects of modifying the design or methods of an experiment. Black students performed differentially well on items measuring the following skills: determine which hypothesis, prediction, or conclusion is, or is not, consistent with two or more theoretical models; determine which additional trial or experiment could be performed to enhance or evaluate experimental results; and make a prediction and explain why it is consistent with two or more theoretical models.

Hispanic-White DIF Analysis

In the Hispanic-White DIF analysis, the available science predictors accounted for 13% of the variance in MH D-DIF. Item discrimination was the most important predictor of MH D-DIF ($r = -.288$, $p < .001$), and it was followed closely by item difficulty ($r = -.216$, $p < .001$). As in the Black-White DIF analysis, more discriminating and easier items favored White students on average. Passage name was the only other predictor with a notable association with MH D-DIF. Overall, passage content and passage standards were not important predictors of DIF, though a few passage standards exhibited systematic DIF. Chemistry topics related to entropy, enthalpy, and calorimetry tended to favor Hispanic students. Earth and space science topics related to the study of the impact of human activity on the Earth including pollution, hazards, remediation, and mitigation; atmospheric structure and conditions; and weather and climate including planets other than Earth tended to favor White students.

Items associated with a small number of skill standards tended to favor Hispanic students: make a prediction and explain why it is consistent with a theoretical model and determine which theoretical models are supported or weakened by new information. The skill standards that favored White students on average related to the following skills: determine how the value of a variable changes as the value of another variable changes in a data presentation, predict the results of an additional trial or measurement in an experiment, and combine data from a data presentation.

Female-Male DIF Analysis

The available science predictors accounted for 15% of the variance in MH D-DIF in the female-male DIF analysis. Item skill standard was the most important predictor of MH D-DIF, followed by passage name and p-value. The item skill standards that favored male and female students the most on average are shown in Table 19. The skill standard that favored males the most was performing an extrapolation using data in a table or graph; the one that favored females the most was identifying similarities and differences between theoretical models. Consistent with these results, items in the reporting category interpretation of data tended to favor males; evaluation of models, inferences, and experimental results items tended to favor females.

Passage name was also strongly related to MH D-DIF. In terms of passage content and passage standards, biology passages, especially those concerning biochemistry and genetics, favored female students on average. Several types of physics passages tended to favor male students: kinematics, electromagnetic waves and optics, and gravity. In addition, Earth and space science passages relating to ocean water and currents, galaxies and the universe, and stars and the solar system tended to favor male students.

Item proportion correct (p-value) was the only quantitative predictor with a notable association with MH D-DIF ($r = .158$, $p < .001$). Easier items tended to favor female students.

Table 19. Science Item Skill Standards That Favored Male and Female Students

Favored male	Favored female
Perform an extrapolation using data in a table or graph (most)	Determine which theoretical models present or imply certain information
Determine and/or use a mathematical relationship that exists between data; e.g., averaging data, unit conversions	Find information in a theoretical model
Identify features of a table, graph, or diagram	Determine which additional trial or experiment could be performed to enhance or evaluate experimental results
Combine data from a data presentation	Determine the scientific question that is the basis for an experiment; e.g., the hypothesis
Make a prediction and explain why it is consistent with two or more theoretical models	Identify similarities and differences between theoretical models (most)

Content Expert Observations

The science content expert team noticed that passages with items that exhibited systematic DIF were often the same across the four DIF analyses. The greatest overlap was between the Black-White and the Hispanic-White DIF analyses. That is, several passages with items that favored White students compared to Hispanic students also favored White students compared to Black students, and several passages that favored Hispanic students compared to White students also favored Black students compared to White students.

For their analysis, the science team calculated the Flesch Reading Ease Score for their passages, and the team found no correlation between reading ease score and MH D-DIF for the Hispanic-White and female-male DIF analyses. However, the science team observed a weak correlation for Black-White DIF analysis, which indicated that relatively easy-to-read passages slightly favored White students compared to Black students.

Analysis of the type and number of data presentations in science passages indicated that the passages with more figures had items that tended to favor White students compared to Black and Hispanic students and also male students compared to female students. Looking at word frequency—particularly how frequently the words *table* and *figure* were mentioned in passages—revealed similar trends.

Importantly, the science team observed that, even though passage-level DIF analyses (i.e., average MH D-DIF statistics for a passage) revealed some tendency for items to favor one group over another, DIF analyses at the individual item level indicated that very few science items exhibited significant DIF. Those that did varied in difficulty, assigned standard, depth of knowledge, and other characteristics.

The science team carefully examined biology units to identify potential explanations for systematic DIF. Out of the eight units selected from the Black-White DIF analysis, the five that favored White students had more complex passages and more challenging items in their item sets. Other than that, there was no clear difference based on the topic chosen, item wording, or difficulty of background knowledge items. A similar trend was observed in the seven units selected from the Hispanic-White DIF analysis. The three units that favored White students had more complex passages and a greater number of challenging items in their item sets. Out of the seven units selected from the female-male DIF analysis, the five that favored female students generally had more complex passages. Note that unit complexity does not necessarily relate to item difficulty, and it is hard to quantify. Many factors, such as the type of data presentation, scientific procedure, the density of text, and familiarity of a topic, contribute to the complexity of passages and items.

Results Summary Tables

Tables 20–23 provide summaries of the DIF trend analysis results reported for the English, math, reading, and science sections of the ACT. The rows of each table list the important predictors of MH D-DIF and the values of those predictors that were most strongly associated with DIF favoring the reference and focal groups. Some important predictors were not included in Tables 20–23 because of their long descriptions. As appropriate, higher-level variables are included (e.g., content standard group or reporting category rather than individual content standard alignment or strand). Otherwise, if a predictor is not listed, then it was not an important predictor or there was no value that systematically favored the reference or focal group.

Table 20. Summary of English Item/Passage Types That Exhibited Systematic DIF

Analysis	Predictor	Favored focal group	Favored reference group
Asian-White	Content standard	Punctuation conventions, sentence structure and formation	Topic development in terms of purpose and focus
Black-White	Difficulty	More difficult	—
	Discrimination	Less discriminating	More discriminating
	Position	Earlier	Later
	Key	—	B (often longest response option)
	Reporting category	—	Production of writing*
	Content standard	—	Topic development in terms of purpose and focus; organization, unity, and cohesion*; style*
	Passage subtype	Language	Chemistry, friends, sociology, literary criticism
Hispanic-White	Position	Earlier	Later
	Content standard	—	Topic development in terms of purpose and focus
	Passage subtype	—	Botany, astronomy
Female-Male	Key	D (last, often shortest response option)	A (most often NO CHANGE)
	Content standard	Sentence structure and formation	Expressing ideas clearly
	Passage subtype	Dance, other personal narrative	Architecture

*A systematic relationship between the predictor and MH D-DIF was observed, but that relationship was substantially weakened after controlling for difficulty and discrimination.

Table 21. Summary of Math Item/Passage Types That Exhibited Systematic DIF

Analysis	Predictor	Favored focal group	Favored reference group
Asian-White	Reporting category	Number and quantity, algebra, functions	Integrating essential skills, statistics & probability
	Real-world context	No	Yes
	Item character count	Lower	Higher
Black-White	Difficulty	More difficult	Easier
	Discrimination	Lower	Higher
	Position	Later	Earlier
	Key	—	E
	Reporting category	—	Integrating essential skills, statistics & probability
	Real-world context	No	Yes
	Item character count	Lower	Higher
Hispanic-White	Difficulty	More difficult	Easier
	Discrimination	Lower	Higher
	Position	Later	Earlier
	Key	—	E
	Reporting category	Algebra	Integrating essential skills, statistics & probability
	Real-world context	—	Yes
	Item character count	Lower	Higher
Female-Male	Reporting category	Algebra, number and quantity	Integrating essential skills, statistics & probability
	Real-world context	—	Yes
	Item character count	Lower	Higher

Table 22. Summary of Reading Item/Passage Types That Exhibited Systematic DIF

Analysis	Predictor	Favored focal group	Favored reference group
Asian-White	—	—	—
Black-White	Discrimination	Lower	Higher
	Position	Earlier	Later
	Content standard	—	Word meanings and word choice, visual and quantitative information
	Passage subtype	Anthropology, short story, environmentalism	Business, astronomy, biology, zoology
Hispanic-White	Discrimination	Lower	Higher
	Position	Earlier	Later
	Content standard	—	Word meaning and word choice
	Passage subtype	Psychology	Ecology, zoology, natural history
	Item character count	—	Low
Female-Male	Difficulty	Easier	More difficult
	Content standard	—	Visual and quantitative information, word meaning and word choice, arguments
	Passage type	Literary narrative	Natural science, social science
	Passage subtype	Personal essay, novel, short story, memoir/autobiography, theater, film, literary criticism	Physics, ecology, astronomy, psychology, technology, architecture, natural science
	Gender representation	Female	—

Table 23. Summary of Science Item/Passage Types That Exhibited Systematic DIF

Analysis	Predictor	Favored focal group	Favored reference group
Asian-White	—	—	—
Black-White	Difficulty	More difficult	Easier
	Discrimination	Lower	Higher
	Reporting category	—	Scientific investigation*
	Passage skill standard	—	Dynamics, electric and magnetic fields
	Passage content	—	Physics
Hispanic-White	Difficulty	More difficult	Easier
	Discrimination	Lower	Higher
	Passage skill standard	Entropy, enthalpy, and calorimetry	Impact of human activity on the Earth including pollution, hazards, remediation, and mitigation; topics related to atmospheric structure and conditions as well as weather and climate including planets other than Earth
Female-Male	Reporting category	Evaluation of models, inferences, and experimental results	Interpretation of data
	Passage skill standard	Biochemistry, genetics	Kinematics, electromagnetic waves and optics, gravity
	Passage content	Biology	Physics, Earth and space science

*A systematic relationship between the predictor and MH D-DIF was observed, but that relationship was substantially weakened after controlling for difficulty and discrimination.

Discussion and Conclusions

The available content, psychometric, and context variables accounted for substantial variation in MH D-DIF statistics, especially for the ACT math section. Consistent with prior research, more difficult items favored Black and Hispanic students, and easier items favored White students. Moreover, female students performed differentially well when reading passages included female representation or were literary narrative texts, and male students performed differentially well when reading sections had science content. Also, as in prior research, White and male students performed differentially well on math word problems with real-world contexts—particularly items assessing statistics and probability or integrating essential skills—whereas focal groups often performed differentially well on algebra and number and quantity items.

Prior studies did not include items measuring science knowledge and skills. Results from this study indicated that female students performed differentially well when science passages featured biology content; White and male students performed differentially well on physics content. As for new variables analyzed in this study, for the English and reading sections, Black and Hispanic students performed differentially well on earlier items, and White students performed differentially well on later items. On the English section, items requiring knowledge of idiomatic language were differentially easy for White students compared to Asian, Black, and Hispanic students. Likewise, reading items that required students to determine meaning from context or the meaning of figurative language were differentially easy for White students. Notably, a construct-irrelevant factor, key, was associated with DIF in several analyses.

This study had several notable limitations. In prior research, differences in language proficiency explained some of the observed DIF between racial/ethnic groups. When students register for the ACT, they can report their first language (English, other, or English and other), but these data are incomplete, and the responses do not indicate current English-language proficiency. For those reasons, this information was not incorporated into the analysis, so some results could be related to English-language proficiency rather than to some other characteristic of racial/ethnic groups. In general, grouping according to broad racial/ethnic groups ignores significant heterogeneity within those groups, including differences in other student characteristics associated with test performance (e.g., eligibility for free or reduced-price lunch). When enough data are available, it would be worthwhile to consider fitting interaction DIF models to account for multiple grouping variables simultaneously (e.g., Finch, 2005). Spurious DIF is another limitation of DIF analyses in general. Specifically, DIF may be associated with item discrimination due to inadequate sum score matching (e.g., matching examinees using number correct, as is typically the case in Mantel-Haenszel DIF analyses), and DIF may be associated with difficulty because of interacting cognitive processes underlying item response behavior (Bolt & Liao, 2021). This report describes observed trends in DIF—similar to those that would be observed in an operational DIF analysis. Attempts were made to adjust for the possible effects of item discrimination and difficulty, but it remains possible that there are genuine (non-spurious) explanations for correlations between DIF and item difficulty and discrimination.

Even if genuine, systematic DIF does not necessarily make a test unfair. For example, female and male students may differ in their interests and preferences for reading passage topics, and that may lead to DIF, but all students should still be expected to demonstrate reading-comprehension skills on literary narrative passages and natural science passages. Moreover, systematic DIF can be explained by other factors such as different patterns in course selection in high school or differential opportunity to learn, which can lead to relative strengths and weaknesses for different student groups on certain types of items. Thus, DIF trends might be examined as guides for addressing systematic differences in achievement within content domains. It may also reflect systematic differences in other student characteristics or behaviors (e.g., motivation, guessing, or omitting).

Future research at ACT could involve matching on IRT ability estimates, which should minimize problems associated with suboptimal sum-score matching. Further studies might investigate possible explanations for DIF trends—especially for construct-irrelevant factors such as key. For example, why is item position associated with DIF (e.g., differential speededness, motivation, omitting/guessing behavior)? With appropriate data, DIF based on other grouping variables may also be informative (e.g., low/high motivation, low/high anxiety, eligibility for free or reduced-price lunch). As more assessments shift to electronic modes of administration, additional data (e.g., click data and response latencies) might be incorporated into DIF research to better understand testing behaviors and psychological states that could manifest as DIF. Finally, results might be used to inform future item development and test blueprints. For example, gender DIF might be preemptively “neutralized” by developing ELA passages about science topics (favors male) with female representation (favors female).

As the ACT North Star asserts, “We exist to fight for fairness in education and create a world where everyone can discover and fulfill their potential” (ACT, 2021, p. 2). The study reported here is part of a larger research agenda to help ensure that tests provide all students with rich opportunities to demonstrate the extent of their mastery of assessed knowledge and skills. Other projects include the development and evaluation of culturally relevant math and science items (Steedle et al., 2023), research on ACT score gains for students with disabilities who test with accommodations (Moore & Schneiders, 2023), and invited panel discussions to support authentic cultural representation in English and reading passages. Please visit *https://www.act.org/content/act/en/research/reports/act-publications.html* to find all our published research reports.

References

ACT. (2021). *2021 year in review: Bringing our north star to life*. *https://www.act.org/content/dam/act/unsecured/documents/2021/2021-Retrospective-Bringing-Our-North-Star-to-Life.pdf*

ACT. (2022). *The ACT® technical manual*. *http://www.act.org/content/dam/act/unsecured/documents/ACT_Technical_Manual.pdf*

Bleistein, C. A., & Wright, D. (1987). Assessment of unexpected differential item difficulty for Asian-American candidates on the Scholastic Aptitude Test. In A. P. Schmitt & N. J. Dorans (Eds.), *Differential item functioning on the Scholastic Aptitude Test* (Research Memorandum No. 87-1). Educational Testing Service.

Bolt, D. M., & Liao, X. (2021). On the positive correlation between DIF and difficulty: A new theory on the correlation as methodological artifact. *Journal of Educational Measurement*, *58*(4), 465–491. *https://doi.org/10.1111/jedm.12302*

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. *https://doi.org/10.1023/A:1010933404324*

Burton, E., & Burton, N. W. (1993). The effect of item screening on test scores and test characteristics. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 321–335). Lawrence Erlbaum.

Carlton, S. T., & Harris, A. M. (1992). *Characteristics associated with differential item functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons* (RR-92-64). Educational Testing Service. *https://files.eric.ed.gov/fulltext/ED385574.pdf*

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*(4), 278–295. *https://doi.org/10.1177/0146621605275728*

Freedle, R., & Kostin, I. (1988). *Relationship between item characteristics and an index of differential item functioning (DIF) for the four GRE verbal item types* (ETS Research Report 88-29). Educational Testing Service. *https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2330-8516.1988.tb00285.x*

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. (1991). *Semantic and structural factors affecting the performance of matched Black and White examinees on analogy items from the Scholastic Aptitude Test* (ETS Research Report Series RR-91-28). Educational Testing Service. *https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.1991.tb01395.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* (Research Report No. 87-48). Educational Testing Service. *https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2330-8516.1987.tb00252.x*

Holland, P. W., & Thayer, D. T. (1986). *Differential item performance and the Mantel-Haenszel procedure* (Research Report No. 86-31). Educational Testing Service. *https://onlinelibrary.wiley.com/doi/10.1002/j.2330-8516.1986.tb00186.x*

Kulick, E., & Dorans, N. J. (1983). *Assessing unexpected differential item performance of Oriental candidates on SAT form CSA6 and TSWE form E33* (Statistical Report No. 83-106). Educational Testing Service.

Kulick, E., & Hu, P. G. (1989). *Examining the relationship between differential item functioning and item difficulty* (Research Report No. 89-18). Educational Testing Service. *https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2330-8516.1989.tb00344.x*

Moore, J. L., & Schneiders, J. Z. (2023). *Score gains of students with disabilities testing with accommodations on the ACT* (Research Report 2023-02). ACT. *https://www.act.org/content/dam/act/unsecured/documents/R2255-Score-Gains-Students-Disabilities-Accommodations-02-2023.pdf*

Randall, J. (2021). “Color-neutral” is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice*, *40*(4), 82–90. *https://doi.org/10.1111/emip.12429*

Rogers, H. J., & Kulick, E. (1987). An investigation of unexpected differences in item performance between Blacks and Whites taking the SAT. In A. P. Schmitt & N. J. Dorans (Eds.), *Differential item functioning on the Scholastic Aptitude Test* (Research Memorandum No. 87-1). Educational Testing Service.

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*

Schmitt, A. P., & Bleistein, C. A. (1987). *Factors affecting differential item functioning for Black examinees on Scholastic Aptitude Test analogy items* (Research Report No. 87-23). Educational Testing Service. *https://files.eric.ed.gov/fulltext/ED292849.pdf*

Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics*, *9*(2), 93–128. *https://doi.org/10.3102/10769986009002093*

Steedle, J. T., Anguiano-Carrasco, C., Lewin, N., & McVey, J. (2023, March 28–30). *Developing culturally relevant math and science items: Lessons learned and student reactions* [Virtual training session]. Annual Meeting of the National Council on Measurement in Education.

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (Research Report No. 12-08). Educational Testing Service. *https://files.eric.ed.gov/fulltext/EJ1109842.pdf*

Appendix

Table A1. English Predictor Variables and Descriptions

Predictor	Description
Difficulty (p-value)	Item difficulty (item proportion correct)
Discrimination (point-biserial correlation)	Item discrimination (item-total correlation)
Position	Item location in the test (1–75)
Key	Correct response (A, B, C, D)
Reporting category	Conventions of standard English (CSE), knowledge of language (KLA), and production of writing (POW)
Depth of knowledge	Webb's DOK level (1–3)
Content standard	Common Core State Standards alignment
Passage name	Unique passage identifier
Passage position	Position of passage within test (1–5)
Passage type	Humanities (HUM), natural science (NSC), social science (SSC), personal narrative (PRL), recreation (REC)
Passage subtype	Various
Gender representation	Gender representation (female, male, both, none)
Ethnicity representation	Ethnic representation (African American/Black [non-Hispanic], American Indian/Alaska Native, Asian American/Pacific Islander, International, Mexican American/Chicano/Latino, Puerto Rican/Cuban/other Hispanic, other, two or more [non-White], none)
Region representation	Regional representation (African, Asian, Central/South American, European, Middle Eastern, North American, other, none)
Urbanicity representation	Geographic representation (urban, rural, none)
Item character count	Number of characters in the item stem and responses
Passage character count	Number of characters in the passage

Table A2. Math Predictor Variables and Descriptions

Predictor	Description
Difficulty (p-value)	Item difficulty (item proportion correct)
Discrimination (point-biserial correlation)	Item discrimination (item-total correlation)
Position	Item location in the test (1–60)
Key	Correct response (A, B, C, D, E)
Reporting category	Number & quantity, algebra, functions, geometry, statistics & probability, integrating essential skills
Depth of knowledge	Webb’s DOK level (1–3)
Strand/topic/standard	Common Core State Standards strand, topic, & standard
Advanced standard	Common Core State Standards indicator of advanced standard, typically taught in 12th grade (no, yes)
Real-world context	Indicator that the item has a real-world context
Modeling	Indicator that the item requires modeling skills
Item character count	Number of characters in the item stem and responses

Table A3. Reading Predictor Variables and Descriptions

Predictor	Description
Difficulty (p-value)	Item difficulty (item proportion correct)
Discrimination (point-biserial correlation)	Item discrimination (item-total correlation)
Position	Item location within test (1–40)
Key	Correct response (A, B, C, D)
Reporting category	Craft & structure (CAS), integration of knowledge & ideas (IKI), key ideas & details (KID)
Understanding complex texts	Indicator that the item requires understanding of central meaning of the text at a level associated with success in college courses with high reading demand (no, yes)
Depth of knowledge	Webb's DOK level (1–3)
Content standard	Common Core State Standards alignment
Passage name	Unique passage identifier
Passage position	Position of passage within test (1–4)
Passage type	Humanities (HUM), literary narrative (LN), natural science (NSC), social science (SSC)
Passage subtype	Various
Gender representation	Gender representation (female, male, both, none)
Ethnicity representation	Ethnicity representation (African-American/Black [non-Hispanic], American Indian/Alaska Native, Asian-American/Pacific Islander, International, Mexican American/Chicano/Latino, Puerto Rican/Cuban/other Hispanic, other, two or more [non-White], none)
Region representation	Regional representation (African, Asian, Central/South American, European, Middle Eastern, North American, other, none)
Urbanicity representation	Geographic representation (urban, rural, none)
Item character count	Number of characters in the item stem and responses
Passage character count	Number of characters in the passage

Table A4. Science Predictor Variables and Descriptions

Predictor	Description
Difficulty (p-value)	Item difficulty (item proportion correct)
Discrimination (point-biserial correlation)	Item discrimination (item-total correlation)
Position	Item location within test (1–40)
Key	Correct response (A, B, C, D)
Reporting category	Evaluation of models, inferences, & experimental results (EMI), interpretation of data (IOD), scientific investigation (SIN)
Depth of knowledge	Webb's DOK level (1–3)
Skill standard	Item alignment to skill standards (various)
Passage skill standard	Passage alignment to content standards (e.g., biology-life science, biology-ecology, chemistry-structure & bonding, Earth & space science-weather & climate, physics-energy)
Passage name	Unique passage identifier
Passage position	Position of passage within test (1–6)
Passage content	Biology (BIO), chemistry (CHE), Earth & space science (ESS), physics (PHY)
Format	Conflicting viewpoints (CV), data representation (DR), research summaries (RS)
Background knowledge	Indicator that the item requires science background knowledge (no, yes)
Background knowledge type	Specific type of science background knowledge required (same values as passage standard)
Item character count	Number of characters in the item stem and responses
Passage character count	Number of characters in the passage

ABOUT ACT

ACT is a mission-driven, nonprofit organization dedicated to helping people achieve education and workplace success. Grounded in more than 60 years of research, ACT is a trusted leader in college and career readiness solutions. Each year, ACT serves millions of students, job seekers, schools, government agencies, and employers in the U.S. and around the world with learning resources, assessments, research, and credentials designed to help them succeed from elementary school through career.

For more information, visit act.org