JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education
DOI: 10.1111/jedm.12372 *Journal of Educational Measurement*

Gender Bias in Test Item Formats: Evidence from PISA 2009, 2012, and 2015 Math and Reading Tests

Benjamin R. Shear
University of Colorado Boulder

Large-scale standardized tests are regularly used to measure student achievement overall and for student subgroups. These uses assume tests provide comparable measures of outcomes across student subgroups, but prior research suggests score comparisons across gender groups may be complicated by the type of test items used. This paper presents evidence that among nationally representative samples of 15-year-olds in the United States participating in the 2009, 2012, and 2015 PISA math and reading tests, there are consistent item format by gender differences. On average, male students answer multiple-choice items correctly relatively more often and female students answer constructed-response items correctly relatively more often. These patterns were consistent across 34 additional participating PISA jurisdictions, although the size of the format differences varied and were larger on average in reading than math. The average magnitude of the format differences is not large enough to be flagged in routine differential item functioning analyses intended to detect test bias but is large enough to raise questions about the validity of inferences based on comparisons of scores across gender groups. Researchers and other test users should account for test item format, particularly when comparing scores across gender groups.

Introduction

Standardized tests are often used to compare educational outcomes among distinct subgroups of students as part of school accountability systems, for research purposes, and to monitor educational equity. The validity of these comparisons rests on assumptions that tests provide comparable measures of student learning across student subgroups and that any differences in scores represent differences in what students know and can do. One aspect of current standardized tests that may be a concern, particularly when scores are compared across gender groups, is whether tests use multiple-choice (MC) items or constructed-response (CR) items. MC items require test-takers to select a response whereas CR items require test-takers to construct a response for an open-ended prompt. Prior research suggests that relative to female students, male students tend to earn relatively higher scores on tests using greater proportions of MC items compared to tests with greater proportions of CR items (Reardon et al., 2018; Schwabe et al., 2015; Taylor & Lee, 2012; Willingham & Cole, 1997).[1] These item format differences pose a threat to the validity of intended interpretations about student achievement, but these threats are not well understood.

Understanding gender by item format differences is especially relevant in the United States given the variability of item formats across tests used to shape public discourse and inform high-stakes decisions. The Common Core-aligned assessments adopted by many states in 2015 tend to rely on a substantial proportion of CR

items to assess more complex thinking skills, as do assessments being developed to assess the Next Generation Science Standards. The same is true of the National Assessment of Educational Progress (NAEP), widely used to monitor national trends in U.S. student achievement and facilitate comparisons across states, and international large-scale assessments such as the Programme for International Student Assessment (PISA). Despite the value of CR items for assessing deeper student learning and the use of CR items in these assessments, standardized tests comprising primarily or entirely MC items also remain common. The SAT and ACT college admissions tests both rely exclusively on MC items (with essay components optional) and are now used in many states for high school accountability testing (Camara et al., 2019; Gewertz, nd). Commercially available interim tests such as the NWEA MAP tests (https://www.nwea.org/the-map-suite/), which are regularly used in classrooms and have been widely cited in education policy discussions, also rely exclusively or primarily on MC items. This variability in item formats raises the possibility that differences in results across gender groups or across tests may be due to the types of items used on each test rather than actual differences in student learning outcomes.

To explore gender by item format differences, this study applies differential item functioning (DIF; Holland & Wainer, 1993) analyses to item response data from nationally representative samples high school-aged students participating in the 2009, 2012, and 2015 PISA Mathematics and Reading Literacy tests (https://www.oecd.org/pisa/). Primary analyses focus on U.S. students participating in PISA and results are then compared across 34 additional participating international jurisdictions. The next section introduces relevant psychometric concepts and prior evidence of gender by item format differences in math and reading tests. The subsequent sections describe the data, methods, and results. The final section summarizes results and discusses implications for the use and development of standardized tests.

Background

The most recent national evidence about gender item format effects in the United States comes from analyses of 2009 state accountability tests in 4th and 8th grade (Reardon et al., 2018). Reardon et al. found that in states using tests with a higher proportion of MC items relative to CR items, male students earned higher scores relative to female students than they did in states using tests with larger proportions of CR items, after adjusting for expected gender differences based on a common test. Reardon et al. conclude that differences in average scores between male and female students would be approximately .10 standard deviations (*SD*s) more male-favoring on tests with 100% MC items compared to tests with 50% MC items, with larger differences in English Language Arts (ELA) than in math, and larger differences in grade 8 ELA than in grade 4 ELA.

To put these results in context, U.S. female students scored approximately .18 to .30 *SD*s higher than male students on the 4th- and 8th-grade NAEP Reading assessments from 2009 to 2019, while male students scored approximately .02 to .06 *SD*s higher than female students on 4th- and 8th-grade NAEP Mathematics assessments.[2] Internationally, among OECD countries participating in PISA 2015 female students scored about .34 *SD* higher than male students in reading, while male students scored

Table 1
Standardized Mean Differences between Male and Female Students across Tests in CO and CT

State	Grade	Year	Test/format	Standardized Mean Difference	
				ELA/EBRW	Math
CO					
	8	2018	CMAS/Mixed	.52	.09
	9	2019	PSAT/MC	.27	.04
CT					
	11	2015	SBAC/Mixed	.32	.14
	11	2016	SAT/MC	.10	–.03

Note. ELA = English Language Arts; EBRW = Evidence-Based Reading and Writing; CMAS = Colorado Measures of Academic Success; SBAC = Smarter Balanced Assessment Consortium. Standardized mean differences in scores represent the difference in means divided by the average within-group standard deviations: $d = (m_f - m_m)/(.5 \times (sd_f + sd_m))$. Positive standardized mean differences indicate female students scored higher than male students.

approximately .08 *SD* higher than female students in math (OECD, 2016). Changing any of these differences by .10 *SD*s would substantially change our inferences about the relative achievement of male and female students.

To provide additional evidence about the potential consequences of gender by item format differences Table 1 reports new analyses for this paper. The analyses use publicly reported aggregate test score data from two states that transitioned from mixed-format tests to MC tests for high school accountability testing. In the first example, 8th-grade Colorado students in 2018 took the Colorado Measures of Academic Success (CMAS) tests in math and ELA, which use both MC and CR items, while in 9th grade in 2019 these students took the PSAT, which uses only MC items.[3] In the second example, 11th-grade Connecticut students in 2015 took the Smarter Balanced (SBAC) tests in math and ELA, which use both MC and CR items, while 11th-grade students in 2016 took the SAT, which uses only MC items. In both states, the standardized mean differences in scores between male and female students were substantially smaller (more male-favoring) for the MC tests relative to mixed-format tests. In ELA standardized mean differences were about .22 to .25 *SD*s more male-favoring on the MC tests and in math the differences were about .06 to .17 *SD*s more male-favoring on the MC tests. Details of these analyses and data are provided in the Supplementary Material.

These cross-test and cross-sample comparisons provide important evidence about the practical consequences of item format effects but face methodological limitations. The comparisons in Table 1 do not provide direct evidence of item format effects because there was no common test administered to adjust for true changes in relative achievement or test participation across years, although it seems reasonable to assume that statewide student cohorts remained similar across years. The examples in Table 1 were included because the observed differences are both substantial and consistent with the presence of previously reported item format effects. The analyses

by Reardon et al. (2018) used a common test to adjust for differences, but the adjustments rely on untestable assumptions about the comparability of scores between the state tests and common test. More direct evidence about item format effects comes from studies that compare the relative performance of a common sample of students on different item types. These studies generally use two different approaches to study item format effects.

The first strategy relies on the use of differential item functioning (DIF) analyses (e.g., Lyons-Thomas et al., 2014; Routitsky & Turner, 2003; Schwabe et al., 2015; Taylor & Lee, 2012). Taylor and Lee (2012), for example, found that among reading and mathematics items flagged for DIF on 1997–2001 Washington state assessments, most items favoring female students were CR and most items favoring male students were MC. The authors report that math items favoring males tended to assess conceptual or procedural understanding, while those favoring females represented a range of mathematics content that included reasoning, problem-solving, and graphing or multiple representations. Among flagged reading items, items favoring male students were more often about identifying reasonable conclusions and interpretations whereas items favoring female students often required students to develop conclusions and analyses. The content of the reading passages and context in which math items were placed did not appear relevant in either domain.

Prior DIF analyses of PISA 2009 data have also found evidence of item format effects. DIF analyses of PISA 2009 mathematics data across four jurisdictions (not including the United States) found that MC items tended to favor male students while CR items tended to favor female students (Lyons-Thomas et al., 2014). A DIF analysis of PISA 2009 reading data in Germany found that, relative to male students, female students earned higher scores on CR items relative to MC items, and this difference remained after controlling for gender differences in reading motivation (Schwabe et al., 2015). These prior studies did not investigate whether systematic differences in other item characteristics between CR and MC items could explain the observed format differences.

The second strategy used to investigate format effects from mixed-format tests computes separate MC and CR scores for each student and then compares relative differences in these separate scores (e.g., DeMars, 2000; Lafontaine & Monseur, 2009; Liu & Wilson, 2009b). An international analysis of PISA 2000 reading results, for example, found a male advantage on MC relative to CR items (Lafontaine & Monseur, 2009). The authors report that the magnitude of the format differences varied depending upon the type of reading process assessed (with slightly larger format differences for items assessing more complex reading processes), but not based on the text type included in the items. In analyses of PISA 2000 and 2003 mathematics data for U.S. and Hong Kong students, Liu and Wilson (Liu & Wilson, 2009a, 2009b) report that score differences between male and female students varied depending on item format and math content assessed, but the patterns were inconsistent across the two countries. Liu and Wilson report that among U.S. students, differences were smallest (i.e., least male-favoring) for standard MC items relative to other item formats, while the differences were largest (most male-favoring) for MC items among Hong Kong students. These studies did not systematically investigate whether item format differences were explained by math content differences.

Although DIF analyses and comparison of MC versus CR scores rely on the same information about relative student performance to estimate format effects, the analyses are intended to inform different audiences and decisions. DIF analyses tend to provide detailed information about specific items flagged with the largest differences, but do not necessarily provide insight into policy-relevant questions about how overall inferences might be impacted by item format changes. Conversely, comparisons of MC and CR scores provide insight into the latter question but do not provide information about item-specific features that might explain format effects. Despite differences in methods and focus these studies are consistent with earlier reviews (Ryan & DeMark, 2002; Willingham & Cole, 1997) suggesting female students tend to earn relatively higher test scores when assessed using tests with more CR items, but the magnitude of the differences are inconsistent across subject areas and contexts.

A key question is whether the gender differences observed are due to construct-relevant or construct-irrelevant factors (Messick, 1995). CR items are often included on tests to assess more complex thinking skills and content than can be assessed with MC items. If CR and MC items assess systematically different aspects of the relevant content domains, differences in how male and female students respond to the items could represent true differences in knowledge or ability. In this case, differential scores by format do not necessarily undermine the validity of inferences based on the resulting scores, although they raise questions about differential learning opportunities that cause such differences. On the other hand, if male and female students respond differently to CR and MC items in ways unrelated to their overall proficiency in the relevant content domain, then differences in scores would reflect construct-irrelevant variance, raising fairness and validity concerns.

There is no consensus explanation for the patterns of observed gender by item format differences that provides an answer to this question. Taylor and Lee (2012) identified construct-relevant factors that differed across items favoring male versus female students. However, these factors were not entirely consistent with prior research about gender differences and the authors did not directly test whether these factors explained the systematic item format differences. In smaller studies analyzing students' written responses in math and reading, female students' higher scores on CR items were attributed partly to female students providing longer answers with more detail, even when these details did not lead to more correct responses (Lane et al., 1996; Pomplun & Capps, 1999; Pomplun & Sundbye, 1999). Others have hypothesized that male students may be more likely to guess on MC items while female students are more likely to skip items (Ben-Shakhar & Sinai, 1991; von Schrader & Ansley, 2006), but differences in response tendencies in these studies did not explain overall differences in performance on MC items.

This study uses a DIF framework to addresses the following three research questions:

1. Is there evidence of differential performance on MC and CR items between male and female students in the 2009, 2012, and 2015 national samples of U.S. 15-year-olds participating in PISA reading and mathematics tests?

2. Are gender by item format differences associated with other properties of the test items among U.S. students?
3. Are gender by item format differences consistent across other international jurisdictions participating in PISA during these years?

The first question investigates whether there are gender by item format differences among recent, nationally representative samples of U.S. 15-year-olds, thus providing more up to date results relative to many of the studies reviewed above. The second question investigates whether item format effects could be due to construct-relevant factors. Finally, the third question investigates whether any observed gender by item format differences are unique to the U.S. context. This study contributes to the literature by combining DIF analyses granular enough to provide information about specific item features with aggregate results that can inform policy-relevant questions about the impact on overall group score comparisons.

Data

PISA Tests

This study uses public student-level item response data files for students participating in the 2009, 2012, and 2015 administrations of PISA obtained from the OECD PISA Database (https://www.oecd.org/pisa/data/). Over 65 jurisdictions (most jurisdictions represent countries, and the term "countries" will be used for the remainder of the paper to refer to these jurisdictions) participated in each PISA cycle from 2009 to 2015. The PISA assessments are intended to assess student "literacy" in reading, mathematics, and science, where literacy is defined as "the extent to which students can apply the knowledge and skills they have learned and practiced at school when confronted with situations and challenges for which that knowledge may be relevant" (OECD, 2017). PISA uses a matrix-sampling design (OECD, 2012, 2014, 2017) in which each student responds to a small number of the total test items rather than responding to all items. This allows PISA to include a more extensive set of items and item formats in the assessments without making the tests too long for individual students. To best represent the constructs that focus on students' ability to apply their knowledge and skills, PISA uses a mix of MC items and CR items. In 2015, PISA transitioned from paper and pencil tests to computerized tests, but the overall design and purpose of the assessments remained consistent.

Student Samples

PISA aims to select nationally representative samples of 15-year-olds attending educational institutions in the fall for participation. In most countries students are selected using a two-stage stratified sampling design; in the first stage schools that enroll 15-year-old students are sampled from within designated strata, and in the second stage random samples of 15-year-old students are sampled within selected schools. In the United States, approximately 5,000 students from approximately 165 schools participated in PISA each year. For the international comparisons, the sample was limited to countries with at least 2,000 students completing the standard test forms of PISA in math and reading in each cycle from 2009 to 2015, administering

paper tests in 2009/2012 and computerized tests in 2015, and to students who responded to at least 5 test items. This resulted in a sample of 35 countries, including the United States, that represent a diverse set of geographic regions and cultural contexts. The Appendix lists all included countries, with additional sample details in the Supplementary Material. The sample size per country per year ranged from 3,839 to 29,359 although not every student completed tests in both math and reading in any given year. The average age among all students was 15.8 years. Student gender is based on each student's response to the question, "Are you female or male?" that provided the response options "female" and "male." It would be valuable to provide empirical evidence about a more complete range of gender identities, but these data are not available in the PISA database.

Item Format and Content

Information about item format and content were obtained from the PISA technical report appendices. First, each item was classified as being either MC or CR. All MC items were scored automatically, while CR items were scored automatically in some cases (e.g., when the correct response was a single number) or by trained scoring teams for more complex responses. PISA reports subcategories within these item formats but the categories changed across years and the more general MC/CR distinction was used here for consistency over time.

PISA employs common items across years for linking; the analyses in this study treat items separately by year because the focus is comparing performance across genders within years, rather than comparing trends in achievement across years. PISA data files exclude data for a small number of items in specific countries for quality control, and a small number of items that almost no students answered correctly were also removed in the present analyses (see below). There were between 412 and 420 item-by-year observations per country, representing 84 unique math items and 104 unique reading items. Approximately 45% of items were MC.

Each reading item was categorized along two dimensions based on the reading process assessed and the format of the text used in the item. The reading items were coded as assessing students' ability to "integrate and interpret," "reflect and evaluate," or "access and retrieve" textual information, and based on whether text accompanying the item was continuous, noncontinuous, or mixed/had multiple formats. Each math item was categorized along two dimensions based on the mathematics content area and mathematical process assessed. The math items were coded as assessing concepts related to either "change and relationships," "quantity," "space and shape," or "uncertainty and data," and based on whether they assessed students' ability to "employ mathematical concepts, facts, procedures, and reasoning," "formulate situations mathematically," or "interpret, apply, and evaluate mathematical outcomes." The Supplementary Material provides additional details about these item features and report the count of items by format and content features.

Item Response Data

The final analytic sample includes approximately 24.5 million item responses from 680,165 students. A small proportion of item-year observations (approximately 9.6%

Table 2
Summary of U.S. Item and Student Samples, by Year and Subject

	Math			Reading		
	2009	2012	2015	2009	2012	2015
Number of items						
All	34	84	68	101	44	88
CR	18	51	39	53	24	46
MC	16	33	29	48	20	42
Average *p* values						
All	.44	.44	.42	.59	.59	.59
CR	.37	.38	.37	.60	.63	.60
MC	.52	.54	.48	.58	.54	.59
Standardized mean differences						
All	–.17	–.06	–.13	.19	.21	.18
CR	–.10	–.01	–.10	.21	.25	.19
MC	–.19	–.10	–.15	.14	.12	.13
Students						
N	3640	4951	2325	5231	3399	2308
% Female	49.2%	49.3%	50.6%	48.7%	49.0%	49.3%

Note. MC = multiple-choice; CR = constructed-response. Item means report the average item percent correct scores. Standardized mean differences in percent correct scores were calculated as the difference in average percent correct scores divided by the average within-group standard deviations: $d = (m_f - m_m)/(.5 \times (sd_f + sd_m))$. Positive values of standardized mean differences indicate females scored higher than males.

for math and 6.4% for reading) were scored polytomously with possible scores of 0, 1, or 2. For the present analyses, all items were converted to binary scores where 0 indicates an incorrect response and 1 represents full or partial credit.[4] Items that were not administered to a student due to the matrix sampling were treated as missing by design. Items that a student skipped (8.2% of all item responses) or did not reach (1.4% of all item responses) were coded as an incorrect response. Within countries and years, any item with a proportion correct less than .01 were dropped (9 of 14,660 country-by-item observations were removed for this reason).

Descriptive Statistics

Table 2 presents descriptive statistics for the U.S. samples, including the total number of items by format. The average percent correct score ("*p* value") by item type, year, and subject indicate two patterns. First, the mathematics items were more difficult on average than the reading items for U.S. students. Second, while the CR math items were more difficult than the MC math items on average, the CR reading items were either similar in difficulty to the MC reading items or easier on average.

Table 2 also reports standardized female-male mean differences in student percent correct scores by item format. The standardized mean differences in percent correct scores using all items shows that male students tended to answer more mathematics items correctly on average in all years, while female students tended to answer

more reading items correctly on average in all years. When calculating the standardized mean differences using only responses to CR or MC items the direction of mean differences remained constant across item formats within subjects. However, the standardized mean differences were consistently more male-favoring when calculated using MC items rather than CR items, providing preliminary evidence of the hypothesized gender by item format effects. The Supplementary Material presents an analogous table summarizing responses in the international samples. Similar patterns were observed for the full international sample.

Methods

A series of item response theory (IRT) models were used to investigate the gender by item format differences more systematically. Many standard DIF analysis methods, such as the Mantel-Haenszel (Holland & Thayer, 1988) or logistic regression (Swaminathan & Rogers, 1990) approaches cannot be readily applied with the matrix-sampling design used in PISA. Hence, a Rasch IRT framework was used to investigate item format effects. A Rasch Model parameterization was selected because this approach was used operationally for the PISA tests in 2009 and 2012 and provides a parsimonious way to summarize the data. Operational PISA scaling used a different IRT model beginning in 2015; the Rasch Model was used for all years in the current analyses for consistency. To address research question (RQ) 1, uniform DIF was estimated for each item in an exploratory analysis using a Rasch Model and then a linear logistic test model (LLTM) was used to test for differential facet functioning (DFF) treating item format as an item facet. To address RQ 2, a series of regression models were used to analyze the item-specific DIF estimates. To address RQ 3, LLTM DFF models were estimated and summarized across countries. All analyses were carried out using the R statistical software (R Core Team, 2022). Computer code to reproduce results are available on GitHub.[5]

To estimate item-specific DIF for each item in the U.S. samples, a many-facets Rasch Model was used to model the log-odds of a correct response to item i by student p as

$$\ln\left(\frac{P\left(X_{ip}=1\right)}{1-P\left(X_{ip}=1\right)}\right)=\theta_p+\gamma F_p-\xi_i+\zeta_i F_p, \tag{1}$$

with $\theta_p \sim N(0, \sigma^2_\theta)$. The variable F_p is an indicator equal to -1 if a student is male and 1 if they are female. The parameter θ_p is a latent variable representing each student's math or reading skill operationalized relative to all items on the test. The parameter γ represents (half) the difference in average scores between male and female students and thus adjusts for overall differences as measured by the set of all items. The parameter ξ_i is the average difficulty of each item in the full sample. The ζ_i parameters represent differences in item difficulty for male and female students. DIF was estimated for each item as $\hat{\delta}_i = 2 \times \hat{\zeta}_i$. The scale of the model was identified by setting the mean of θ_p to 0 and constraining the sum of all item DIF estimates to 0, i.e., $\Sigma\zeta_i = 0$. These models were estimated using the function tam.mml.mfr in the R package TAM (Robitzsch et al., 2022) separately for each year and subject.

The $\hat{\delta}_i$ are the primary estimates of interest. Positive values of $\hat{\delta}_i$ indicate that female students were more likely to answer an item correctly, after adjusting for differences in average math or reading scores between male and female students. The reverse is true for negative values of $\hat{\delta}_i$. In many standard DIF analysis contexts the magnitude of each item DIF statistic would be compared to set thresholds for statistical and practical significance in order to identify items displaying DIF (Zumbo, 1999). These items would then be screened to examine potential explanations for the DIF.

To provide a direct test of item format differences, item format was treated as an item facet variable and DFF was tested using a LLTM with random item errors (De Boeck & Wilson, 2004b). The LLTM models item difficulty as a linear combination of one or more item properties rather than estimating a unique difficulty or DIF value for each item, thus providing a more direct and parsimonious test of item format differences. In this model, the log-odds of a correct item response was modeled as

$$\ln\left(\frac{P\left(X_{ip}=1\right)}{1-P\left(X_{ip}=1\right)}\right)=\beta_0+\beta_1 F_p+\beta_2 CR_i+\beta_3\left(F_p\times CR_i\right)+\eta_p+\varepsilon_i, \qquad (2)$$

where $\eta_p \sim N(0, \sigma^2_\eta)$ and $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$ are random effects for students and items, respectively. Here F_p is defined as in Equation 1 and CR_i is an indicator equal to 1 if an item is CR and 0 otherwise. The person random effects represented by η_p can be interpreted equivalently to θ_p in Equation 1. The random item errors represented by ϵ_i were included to account for variation in item difficulty not explained by item format and to maintain a comparable latent scale between the models in Equations 1 and 2 (De Boeck & Wilson, 2004a, pp. 32−33). Under the assumption that the test items measure a unidimensional construct, the parameter β_3 in Equation 2 represents (half) the difference in the difficulty of CR items for female versus male students, relative to the difference in difficulty of MC items for female versus male students. Positive values of β_3 indicate that, relative to male students, female students were more likely to answer CR items correctly, adjusting for differences in ability and likelihood of answering MC items correctly. These models were estimated using the glmer function in the R package lme4 (Bates et al., 2015; De Boeck et al., 2011).

When IRT is used to scale a test, one goal is to estimate the ability of each test-taker on the logit scale, relative to the locations of the items. If two populations of students have the same level of ability relative to the construct being measured, but the items used to scale respondents are easier on average for one population, the average location estimate of respondents will be shifted by an equivalent amount. As a result, we would incorrectly infer a difference in ability. A standardized format effect can be defined as

$$d_{format}=\frac{2\times\hat{\beta}_3}{\sqrt{\hat{\sigma}^2_\eta}}. \qquad (3)$$

This standardized format effect estimates how much the standardized mean difference in scores between male and female students would change if moving from a test with entirely MC items to one with entirely CR items. A positive value indicates that

moving from a test with entirely MC items to entirely CR items would favor female students. Estimates of d_{format} are reported to characterize the effect size of estimated item format differences.

To address RQ 2 least squares multiple linear regression models were used to examine whether differences in DIF between MC and CR items might be due to other item features. The following regression model was estimated separately for each subject for the U.S. item-specific DIF estimates:

$$\hat{\delta}_{iy} = \alpha_0 + \alpha_1 (CR_i) + \omega X_{iy} + \Delta_y + e_{iy}. \quad (4)$$

The $\hat{\delta}_{iy}$ is the estimated DIF coefficient from Equation 1 for item i in year y, CR_i is an indicator equal to 1 for CR items and 0 for MC items, $\boldsymbol{X}_{iy}$ is a vector of item covariates, and Δ_y are year fixed-effects. Item covariates include estimated item difficulty ($\hat{\xi}_i$) and indicators for the math content and process assessed (math items) or the reading process and text format (reading items). The coefficient α_1 represents the unstandardized format effect, conditional on other item covariates included in the model. Models were estimated with and without $\boldsymbol{X}_{iy}$ to examine whether the format effect was still present after adjusting for item covariates. If the difference in DIF for MC and CR items is due to differences in overall item difficulty, content, or process assessed, we would expect α_1 to be smaller when including these item features.[6]

To address RQ 3 the model in Equation 2 was fit separately to data for each subject and year in each of the remaining 34 countries to estimate d_{format} from Equation 3 across countries. Although it would be possible to estimate the model in Equation 2 by pooling across countries (within subjects and years), this was not pursued because PISA tests are administered in different languages across countries. Estimating item format differences in a single model pooling across countries would make additional assumptions about the invariance of item properties and definition of the latent scale across different language versions of the test. Estimating the models separately across countries allows the format effect within each country to be interpreted without potential confounding due to language differences.

Results

The top portion of Table 3 summarizes the average and range of item DIF estimates by year, subject, and item format among U.S. samples. Consistent with the hypothesized format effects, the average DIF for CR items was positive in every administration ($M = .05$ logits overall) and average DIF for MC items was negative in every administration ($M = -.06$ logits overall). Average DIF showed a consistent pattern, but there was substantial variation in DIF estimates within years, subjects, and item formats. The average DIF for MC and CR items was smallest in 2015, the only year in which tests were administered on computers. This provides preliminary evidence that item format effects may differ for computerized versus paper tests, but we should be cautious about generalizing from a single year of data and further investigation is warranted. Overall, 82 of the 186 math DIF estimates (44.1%) were statistically significant at the $p < .05$ level, while 112 of the 233 reading DIF estimates (48.1%) were statistically significant. The proportion of items with significant DIF was similar for MC and CR. Due to the identification constraints of the model,

Table 3
DIF Summary Statistics and Standardized Format Effects, by Year and Subject among U.S. Samples

	Math			Reading		
	2009	2012	2015	2009	2012	2015
Rasch Model DIF						
CR items						
Mean	.04	.06	.01	.05	.08	.04
Min	–.35	–.58	–.45	–.61	–.55	–.80
Max	.40	.88	.65	.49	.42	.61
MC items						
Mean	–.05	–.09	–.01	–.06	–.09	–.04
Min	–.55	–.63	–.47	–.56	–.43	–.60
Max	.37	.49	.49	.32	.57	.83
$sd(\theta)$	1.20	1.29	1.22	1.31	1.23	1.34
d	–.20	–.07	–.16	.20	.27	.21
LLTM DFF						
β_3	.05*	.08***	.02	.05***	.09***	.07**
$se(\beta_3)$	.02	.01	.02	.01	.02	.02
d_{format}	.08	.13	.04	.08	.15	.10
$sd(\eta)$	1.18	1.28	1.20	1.30	1.21	1.32
$sd(\epsilon)$	1.19	1.48	1.44	1.28	1.48	1.37

Note: CR = constructed-response; MC = multiple-choice; d = standardized mean difference described in text (positive values indicate female students score higher); d_{format} = standardized item format effect (positive values indicate a female advantage for CR items). Positive DIF values indicate items easier for female students.
*$p < .05$. **$p < .01$. ***$p < .001$.

approximately equal numbers of items favored female and male students. Table 3 also reports the estimated *SD* of θ and the estimated standardized mean difference of θ across gender estimated as $d = (2 \times \hat{\gamma})/\hat{\sigma}_\theta$ from Equation 1. The estimates of d indicate that male students scored higher on the math tests on average while female students scored higher on the reading tests.

The bottom portion of Table 3 presents estimates and standard errors of β_3, standardized item format effects, and estimated *SD* of the random effects based on the LLTM DFF models among U.S. samples. The estimated *SD* of θ in the Rasch Models from Equation 1 were within .02 of the estimated *SD* of η for the LLTM, indicating the models are estimated on comparable latent scales. Multiplying the estimates of β_3 by two yields similar (although not identical) values to the difference in average DIF between CR and MC items. The estimates of β_3 were positive in every administration and were statistically significant at the $p < .05$ for all administrations except 2015 math. The standardized format effects range from .04 (2015 math) to .15 (2012 reading) with an average of .08 in math and .11 in reading ($M = .10$ overall). This suggests that, on average, moving from a test based entirely on CR items to one based entirely on MC items could change the standardized mean difference in scores

Table 4
Regression Model Estimates Predicting Item DIF for U.S. Sample

	Math 1		Math 2		Reading 1		Reading 2	
Predictors	*Est.*	*SE*	*Est.*	*SE*	*Est.*	*SE*	*Est.*	*SE*
Intercept	–.05	.05	–.01	.07	–.06	.03	–.06	.05
Is CR	.09*	.04	.13**	.04	.11**	.04	.09*	.04
Difficulty			–.02	.01			–.07***	.01
Quantity			.01	.05				
Space and shape			–.08	.05				
Uncertainty and data			–.01	.05				
Formulate			–.11*	.05				
Interpret			–.04	.05				
Integrate and interpret							.02	.05
Reflect and evaluate							.07	.05
Text format mixed/multiple							–.07	.06
Text format noncontinuous							–.05	.04
Observations	186		186		233		233	
R^2	.026		.125		.039		.160	

Notes: All models include year fixed effect indicators. The omitted category for math content is "change and relationships"; the omitted category for math process is "employ;" the omitted category for reading process is "access and retrieve"; the omitted category for text format is "continuous." Item difficulty estimates are mean centered by subject.
*$p < .05$. **$p < .01$. ***$p < .001$.

by .05 to .10 *SD*, with slightly larger differences in reading than in math. In an operational testing context focused on identifying biased test items, a single item with DIF of .05 to .10 logits would generally not be flagged as problematic, even if the DIF were statistically significant. However, the systematic pattern of these results highlights that while this level of DIF may be negligible for a single item, it could be practically relevant if it accumulates systematically across many items.

Table 4 presents the regression analysis results for the U.S. sample DIF estimates. In both subjects, the item format differences remained after adjusting for other item covariates. The set of item covariates only explained 12–16% of the variance in item DIF across subjects. Because the item DIF estimates are not independent and contain sampling error, the standard errors and *p* values should be interpreted cautiously. In math, the conditional format difference was slightly larger (.13 versus .09) when controlling for difficulty, math content, and math process and would be statistically significant at the $p < .01$ level. In reading, the conditional format difference was slightly smaller (.09 versus .11) when controlling for difficulty, reading process, and text format, and would be statistically significant at the $p < .05$ level. In math the only additional item covariate with an association that would reach traditional levels of significance was the indicator for formulate, suggesting that math items requiring students to formulate situations mathematically tended to favor male students relative

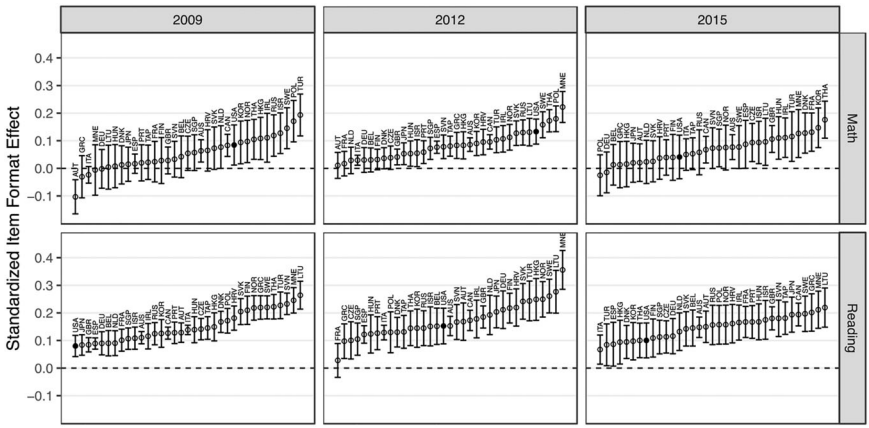

Figure 1. Estimated standardized item format effects across countries, by year and subject.

Note. Error bars denote approximate 95% confidence intervals. Estimates are sorted by magnitude within year and subject. Solid points represent estimates for U.S. samples. Positive values indicate a female advantage on CR items. Full country names are listed in the Appendix.

Table 5

Standardized Item Format Effect Summary Statistics across Countries, by Year and Subject

	Math			Reading		
	2009	2012	2015	2009	2012	2015
Mean	.05	.08	.07	.15	.17	.15
Min	–.10	.01	–.03	.08	.03	.07
Max	.19	.22	.18	.26	.36	.22
N Sig.	10	25	11	35	34	31

Note: Estimates summarized across $N = 35$ countries. Num. Sig. = number of estimates statistically significant at the $p < .01$ level across countries.

to items requiring students to employ mathematics. In reading, the only additional statistically significant coefficient was for item difficulty, suggesting item DIF for more difficult items tended to favor male students. This pattern is consistent with prior studies suggesting male students tend to do relatively better on more difficult items, although prior research has more often focused on gender by item difficulty associations in MC mathematics items (e.g., Bielinski & Davison, 2001).

Figure 1 plots the standardized item format effects based on the LLTM DFF models across all 35 countries and Table 5 summarizes the distribution of these estimates, including the number of β_3 estimates that would be statistically significant at the $p < .01$ level. There is consistent evidence of item format effects across countries showing a female advantage on CR items, with greater differences in reading. The average standardized format effects across countries and years are .07 in math and

.16 in reading. In reading, every estimated item format effect is positive, and the majority are statistically significant. In math, 98 of the 105 total format effect estimates are positive, but many are not statistically significantly different from 0. A small minority of the format effects in math (7 out of 105) are negative, and 1 is statistically significantly less than 0. These results suggest the item format effect is not unique to the United States, although the magnitude of the effects varies across countries. The results do not appear to differ noticeably in 2015 relative to 2009 and 2012.

Discussion

The results of this study provide consistent evidence that among recent samples of high school-aged students participating in PISA, on average male students tend to earn relatively higher scores on MC test items whereas female students tend to earn relatively higher scores on CR test items. Among U.S. students, the magnitude of the standardized format effects was similar across subjects. Internationally the format effects were consistently stronger in reading than in math. The substantial variability in the item-specific DIF among U.S. samples is an important reminder that although format difference were observed consistently across years and subjects, male students do not always do better on all MC items and vice versa for CR items. The variation in the magnitude of the standardized format effects across countries suggests there are relevant social and cultural factors that likely contribute to these differences as well.

The observed format differences in the present analysis were generally smaller than the test-level differences reported in prior research (e.g., Reardon et al., 2018) and reported above in Table 1. The item format effect estimates in the analyses by Reardon et al. (2018) were approximately .20 *SD* on average across grades and subjects. The test-level results presented in Table 1 are also consistent with format differences as large as .25 *SD*. In the PISA analyses, the average standardized item format effects across countries were .07 in math and .16 in reading (.08 and .11, respectively, among U.S. samples), indicating that the standardized mean difference in scores would be .07 to .16 *SD* more male-favoring on a test with entirely MC items relative to one with entirely CR items on average. However, format effects of .07 to .16 *SD* are practically relevant compared to the overall average differences in scores observed between male and female students and to the magnitudes of effect sizes often observed in educational research (e.g., Kraft, 2020). One potential explanation is that the mixed-format and MC tests included in prior analyses measure different knowledge and skills than PISA in addition to using different item formats.

Among U.S. samples, the estimated item format differences were similar after controlling for item difficulty, math content, math process, reading process, and text type of each item. This suggests the format effects were not due primarily to systematic differences between MC and CR items in these additional item features. Although this result is consistent with the hypothesis that construct-irrelevant factors caused the item format differences, no strong causal conclusions are warranted based on these results alone. There was limited information about item content that could be included in the analyses, and the unexplained variability of the DIF

estimates underscores that there are other factors driving both item-specific and overall differences in performance across gender groups.

There are some important limitations to these analyses worth noting. We cannot directly infer that the format differences were caused by the item format, rather than other systematic differences between item types not measured here. From a generalization standpoint, although representative samples of students participate in PISA, sampling weights were not used in the analyses (as is common in other DIF analyses with PISA), suggesting caution generalizing the results to the broader populations of high school students in each country. Caution is also warranted for generalizing to other tests beyond the low-stakes PISA tests. The use of the IRT DIF and DFF models assume that there is a unidimensional construct measured by each test and defined in the aggregate by the combination of MC and CR items. The IRT model adjusts for overall performance relative to this construct, and relative item format differences are also defined relative to the construct. Based on the PISA design process and classical item statistics analyzed, this assumption appears plausible, while also acknowledging the systematic differences in performance suggest potential multidimensionality.

Finally, these analyses compared performance across student gender defined dichotomously (male or female). This was done with the recognition that gender identity is a more complex phenomenon not fully represented by this categorization, and that there is considerable variability within these two groups. This study also did not investigate whether aspects of students' identity or experiences such as their racial/ethnic identity, family economic situation, school environment, or other country-level factors were associated with differential performance. Investigating item format differences across groups defined by other identities, or between intersections of these identities, as well as possible explanations for cross-country variability is an important avenue for future work given the widespread use of test scores to document and study disparities in educational opportunities between groups.

Despite these limitations, the results presented here have important implications for the use and interpretation of large-scale tests. First, more research is needed to identify potential explanations of item format differences and to determine whether they are due to construct-relevant or construct-irrelevant factors. If the differences are caused by construct-irrelevant factors, then these sources need to be identified and test designs may need to be modified to reduce the impact on student scores. If the differences are caused by construct-relevant factors, additional investigation of differential learning opportunities are needed, and attention is warranted to ensure that the constructs represented by a test accurately match the skills and knowledge the test is intended to assess. Because test developers have access to the most detailed item response data, they should be expected to document and explain the presence of item format effects on their tests.

One avenue for investigation is students' level of effort and omission rates across items. Although not part of the planned analysis, post-hoc exploratory analyses revealed that on average male students were more likely to omit item responses and this difference was more pronounced for CR than for MC items and in reading relative to math. This contradicts the hypothesis that males may have an advantage on MC items because they are more willing to guess whereas female students are more

likely to omit a response; differences existed in both subjects despite female students having lower overall scores in math and higher overall scores in reading. This differential pattern of omissions is consistent with recent research reporting that male students were more likely than female students to provide low effort rapid guesses to MC items that required a response in a computer adaptive testing environment (Soland, 2018). However, the format differences in the present analysis do not appear to be solely due to the scoring of omitted responses as incorrect. As a sensitivity analysis, the item format effects were reestimated after replacing omitted responses (including not reached items) with the modal nonmissing response for each item. The substantive conclusions and direction of DFF across item formats remained the same, although the average standardized format effects were smaller in reading and about the same in math. Further exploration of the interaction among response times, item omissions, and item formats could leverage recently developed models that incorporate response times and item omissions (e.g., Debeer et al., 2017; Lu & Wang, 2020).

Second, regardless of the cause of the format differences, those responsible for selecting and using large-scale standardized tests should take item formats into account when selecting tests for different purposes and interpreting the results. Decisions about whether to use mixed-format or entirely MC tests may not have a large effect on inferences for aggregate groups when the proportion of male and female students across groups is relatively similar. However, when tests are used to make decisions about individual students the types of items used could be consequential. If tests are used for course placements or admission into programs, for example, there could be consequences for the proportion of male and female students selected for different opportunities. The differences are also relevant when tests are used as equity indicators to monitor achievement outcomes across student subgroups that include gender groups, and support calls to further investigate causes of observed differences in scores to promote gender equity (Cimpian, 2020).

Third, these analyses highlight that standard practices focused on items with especially large DIF may not be sufficient to address fairness issues related to the use of different item formats. Many of the differences observed between male and female students for individual items were not large enough to be flagged as "problematic" in typical assessment validation processes relying on DIF analysis. It was the compounding nature of the DIF aggregated across items that led to the systematic format differences. Large-scale assessments such as PISA or the state tests used in the test-level analyses cited above regularly include monitoring for DIF as part of the test development and validation process. However, not all possible DIF comparisons are undertaken for every assessment. In PISA, for example, DIF analyses focus primarily on between-country DIF rather than within-country DIF across gender groups. The systematic item format differences found across gender groups is a reminder that threats to the validity of score comparisons can exist even in carefully designed standardized tests that have undergone common psychometric evaluations for bias and fairness.

Acknowledgments

The author acknowledges helpful research assistance from Kyla McClure, Sarah Wellberg, and Medjy Pierre-Louis. This research was supported by a grant from the American Educational Research Association which receives funds for its "AERA Grants Program" from the National Science Foundation under NSF award NSF-DRL #1749275. Opinions reflect those of the author and do not necessarily reflect those of AERA or NSF. An earlier version of this manuscript was presented at the 2022 National Council on Measurement in Education Annual Meeting.

Notes

[1]Students' gender identities are not fully represented by a simple male/female dichotomy. However, this study compares performance using a male/female dichotomy for two reasons. First, state accountability systems and other reports regularly present scores for male/female gender subgroups, and research has found evidence of relevant test score differences across these two groups that needs to be better understood. Second, more detailed information about students' gender identities was not available in the data.

[2]These figures are based on calculations using data from https://nces.ed.gov/nationsreportcard/data/.

[3]All reading and writing items on the PSAT and SAT are MC. Most items on the math test are MC, with a small number of items requiring students to provide a numeric answer by filling in appropriate bubbles.

[4]Binary scoring was used to facilitate the estimation of the models with random item errors described below. Preliminary analyses that estimated item-specific DIF using a partial credit parameterization for polytomous items produced substantively similar results for the analyses based on item-specific DIF, suggesting this analytic decision would not materially affect the conclusions.

[5]Available at https://github.com/bshear/pisa_format_dif.

[6]Random effects meta-analytic regression models accounting for the uncertainty of the DIF estimates were also estimated and results were substantively the same. These results are reported in the Supplementary Material.

References

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement*, *28*(1), 23–35. https://doi.org/10.1111/j.1745-3984.1991.tb00341.x

Bielinski, J., & Davison, M. L. (2001). A sex difference by item difficulty interaction in multiple-choice mathematics items administered to national probability samples. *Journal of Educational Measurement*, *38*(1), 51–77. https://doi.org/10.1111/j.1745-3984.2001.tb01116.x

Camara, W. J., Mattern, K., Croft, M., Vispoel, S., & Nichols, P. (2019). A validity argument in support of the use of college admissions test scores for federal accountability. *Educational Measurement: Issues and Practice*, *38*(4), 12–26. https://doi.org/10.1111/emip.12293

Cimpian, J. R. (2020). Why focusing on test metrics may impede gender equity: Policy insights. *Policy Insights from the Behavioral and Brain Sciences*, *7*(1), 64–71. https://doi.org/10.1177/2372732219873009

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*(12), 1–28.

De Boeck, P., & Wilson, M. (2004a). A framework for item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models* (pp. 3–41). New York: Springer. https://doi.org/10.1007/978-1-4757-3990-9_1

De Boeck, P., & Wilson, M. (Eds.). (2004b). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer. https://doi.org/10.1007/978-1-4757-3990-9

Debeer, D., Janssen, R., & De Boeck, P. (2017). Modeling skipped and not-reached items using IRTrees: Modeling skipped and not-reached items using IRTrees. *Journal of Educational Measurement*, *54*(3), 333–363. https://doi.org/10.1111/jedm.12147

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*(1), 55–77. https://doi.org/10.1207/s15324818ame1301_3

Gewertz, C. (nd). What tests does each state require? *EducationWeek*. https://www.edweek.org/teaching-learning/what-tests-does-each-state-require

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

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

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189x20912798

Lafontaine, D., & Monseur, C. (2009). Gender gap in comparative studies of reading comprehension: To what extent do the test characteristics make a difference? *European Educational Research Journal*, *8*(1), 69–79. https://doi.org/10.2304/eerj.2009.8.1.69

Lane, S., Wang, N., & Magone, M. (1996). Gender-related differential item functioning on a middle-school mathematics performance assessment. *Educational Measurement: Issues and Practice*, *15*(4), 21–27. https://doi.org/10.1111/j.1745-3992.1996.tb00575.x

Liu, O. L., & Wilson, M. (2009a). Gender differences and similarities in PISA 2003 Mathematics: A comparison between the United States and Hong Kong. *International Journal of Testing*, *9*(1), 20–40. https://doi.org/10.1080/15305050902733547

Liu, O. L., & Wilson, M. (2009b). Gender differences in large-scale math assessments: PISA Trend 2000 and 2003. *Applied Measurement in Education*, *22*(2), 164–184. https://doi.org/10.1080/08957340902754635

Lu, J., & Wang, C. (2020). A response time process model for not-reached and omitted items. *Journal of Educational Measurement*, *57*(4), 584–620. https://doi.org/10.1111/jedm.12270

Lyons-Thomas, J., Sandilands, D., & Ercikan, K. (2014). Gender differential item functioning in mathematics in four international jurisdictions. *Education and Science*, *39*(172), 20–32.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

OECD. (2012). *PISA 2009 technical report*. OECD Publishing. https://doi.org/10.1787/9789264167872-en

OECD. (2014). *PISA 2012 technical report*. OECD Publishing. https://www.oecd.org/pisa/pisaproducts/pisa2012technicalreport.htm

OECD. (2016). *PISA 2015 results (Volume I): Excellence and equity in education*. OECD. https://doi.org/10.1787/9789264266490-en

OECD. (2017). *PISA 2015 technical report*. OECD Publishing. https://www.oecd.org/pisa/data/2015-technical-report/

Pomplun, M., & Capps, L. (1999). Gender differences for constructed-response mathematics items. *Educational and Psychological Measurement*, *59*(4), 597–614. https://doi.org/10.1177/00131649921970044

Pomplun, M., & Sundbye, N. (1999). Gender differences in constructed response reading items. *Applied Measurement in Education*, *12*(1), 95–109. https://doi.org/10.1207/s15324818ame1201_6

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A., & Zárate, R. C. (2018). The relationship between test item format and gender achievement gaps on math and ELA tests in fourth and eighth grades. *Educational Researcher*, *47*(5), 1–11. https://doi.org/10.3102/0013189x18762105

Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test Analysis Modules* (R package version 4.1). https://cran.r-project.org/web/packages/TAM/TAM.pdf

Routitsky, A., & Turner, R. (2003). *Item format types and their influence on cross-national comparisons of student performance*. Annual Meeting of the American Educational Research Association.

Ryan, J. M., & DeMark, S. (2002). Variation in achievement scores related to gender, item format, and content area tested. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 57–75). Routledge.

Schwabe, F., McElvany, N., & Trendtel, M. (2015). The school age gender gap in reading achievement: Examining the influences of item format and intrinsic reading motivation. *Reading Research Quarterly*, *50*(2), 219–232. https://doi.org/10.1002/rrq.92

Soland, J. (2018). The achievement gap or the engagement gap? Investigating the sensitivity of gaps estimates to test motivation. *Applied Measurement in Education*, *31*(4), 312–323. https://doi.org/10.1080/08957347.2018.1495213

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

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education*, *25*(3), 246–280. https://doi.org/10.1080/08957347.2012.687650

von Schrader, S., & Ansley, T. (2006). Sex differences in the tendency to omit items on multiple-choice tests: 1980–2000. *Applied Measurement in Education*, *19*(1), 41–65. https://doi.org/10.1207/s15324818ame1901_3

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Lawrence Erlbaum Associates.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-Type (ordinal) item scores*. Directorate of Human Resources Research and Evaluation, Department of National Defense. http://faculty.educ.ubc.ca/zumbo/DIF/

Author

BENJAMIN R. SHEAR is Assistant Professor in the Research and Evaluation Methodology Program at the University of Colorado Boulder School of Education, 249 UCB, Boulder, CO 80309; benjamin.shear@colorado.edu. His research focuses on applied statistical issues in educational measurement and psychometrics, particularly those relevant for validity and validation.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix