


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

Benjamin R. Shear

School of Education, University of Colorado Boulder

Author Note

Benjamin R. Shear  <https://orcid.org/0000-0002-9236-2927>

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.

Correspondence concerning this article should be addressed to Benjamin R. Shear, School of Education, University of Colorado Boulder, 249 UCB, Boulder, CO 80309.

Email: benjamin.shear@colorado.edu

Abstract

Large-scale standardized tests are regularly used to measure trends in student achievement overall and for student subgroups. The results are used for system monitoring, research studies, and school accountability systems. These uses assume tests provide comparable measures of student outcomes across student subgroups, but prior research suggests score comparisons across student gender groups may be complicated by the type of test items used. This paper presents evidence that among recent nationally representative samples of 15-year-olds in the US participating in the 2009, 2012, and 2015 PISA tests, there is consistent evidence of item format by gender differences. Male students answer multiple-choice items correctly relatively more often and female students answer constructed-response items correctly relatively more often, raising questions about the validity of intended inferences. The paper discusses the magnitude of these differences and implications for test use.

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 given the variability of item formats across tests used to shape public discourse and inform high-stakes decisions in the US. 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

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.

of the National Assessment of Educational Progress (NAEP), widely used to monitor national trends in US 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 not to actual differences in student learning outcomes, but instead due to the types of items used on each test.

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 of US high school students participating in the 2009, 2012, and 2015 PISA Mathematics and Reading Literacy tests (<https://www.oecd.org/pisa/>). 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-level evidence about gender item format effects in the US comes from analyses of 2009 state accountability testing in 4th and 8th grade. Analyzing data from nearly all 50 states, researchers 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., 2018). The researchers estimated the difference in average scores between male and female students would be approximately 0.10 standard deviations (SDs) more male-favoring on a test with 100% MC items compared to a test with 50% MC items. The differences were larger in English Language Arts (ELA) than in math, and larger in grade 8 ELA than in grade 4 ELA. To put these results in context, in 4th and 8th grade nationally, female students scored approximately 0.18 to 0.30 SDs higher than male students on the NAEP Reading assessments from 2009-2019, while male students scored approximately 0.02 to 0.06 SDs higher than female students on NAEP Mathematics assessments.² Changing any of these differences by 0.10 SDs would substantially change our inferences about the relative achievement of male and female students.

To provide additional more recent examples, Table 1 summarizes 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, Colorado students in 8th grade took the Colorado Measures of Academic Success (CMAS) tests in math and ELA, which use both MC and CR items; students in 9th grade took the PSAT, which uses only MC items.³ In the

² Author 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.

second example, Connecticut 11th grade students in 2014-15 took the Smarter Balanced (SBAC) tests in math and ELA, which also use a mix of MC and CR items, but in 2015-16 11th grade students took the SAT, which uses MC items. Table 1 presents results for the 2017-18 8th grade cohort in Colorado (who took CMAS tests in 8th grade in 2018 and PSAT tests in 9th grade in 2019) and the 2014-15 and 2015-16 cohorts of 11th grade students in Connecticut (who took the SBAC and SAT tests, respectively). In both states the standardized mean differences in scores were substantially smaller (and thus more male-favoring) when measured using MC tests relative to mixed-format tests.⁴ In ELA, standardized mean differences were about 0.22 to 0.25 SDs more male-favoring on the MC tests, while in math the differences were about 0.06 to 0.17 SDs more male favoring on the MC tests. In Connecticut, the direction of the difference reversed for 11th grade students when switching from the SBAC to SAT test.

Table 1.

Standardized Mean Differences between Male and Female Students Across Tests in CO and CT

State	Grade	Year	Test	Subject	Format	<i>d</i>
CO	8	2018	CMAS	ELA	Mixed	0.52
	9	2019	PSAT	EBRW	MC	0.27
	8	2018	CMAS	Math	Mixed	0.09
	9	2019	PSAT	Math	MC	0.04
CT	11	2015	SBAC	ELA	Mixed	0.32
	11	2016	SAT	EBRW	MC	0.10
	11	2015	SBAT	Math	Mixed	0.14
	11	2016	SAT	Math	MC	-0.03

Note. *d*=standardized mean difference; ELA=English Language Arts; EBRW=Evidence-Based Reading and Writing; CMAS=Colorado Measures of Academic Success; SBAC=Smarter Balanced Assessment Consortium. Positive standardized mean differences indicate females scored higher than males.

⁴ Differences for CO were estimated based on publicly reported scale score means and standard deviations. Differences for CT were estimated via heteroskedastic ordered probit models based on publicly reported proficiency level counts using the -hetop- package in Stata (Shear & Reardon, 2019).

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 participation across years, although it seems reasonable to assume that statewide student cohorts remained similar across years. These examples are 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) did use 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). DIF analysis is a psychometric technique developed to identify potentially biased test items by estimating between-group differences in performance on specific items after adjusting for overall performance across groups (Zumbo, 1999). Standard DIF analyses begin by flagging items with significant group differences in item success rates. In a second step, researchers or test developers examine item characteristics and attempt to explain the observed differences to determine whether observed DIF is due to construct relevant or construct irrelevant factors. Numerous DIF analyses of mixed-format tests find that items flagged as favoring female students are disproportionately CR items, whereas items favoring male students tend to be MC

items, although they have generally not been able to identify specific explanations for this pattern.

Taylor and Lee (2012), for example, found that among reading and mathematics items flagged for DIF on 1997-2001 Washington state assessments, most of the items favoring female students were CR and most of the 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 in other countries have also found evidence of item format effects. DIF analyses of PISA 2009 mathematics data (Lyons-Thomas et al., 2014) found that MC items tended to favor male students while CR items tended to favor female students although US students were not included in the study. A study of PISA 2009 reading data in Germany used DIF analysis to compare across all items, and found that after adjusting for overall reading ability female students earned higher scores on CR items relative to MC items (Schwabe et al., 2015). Among 15-year-olds this difference remained after adjusting for differences in students' intrinsic reading motivation, thus ruling out the authors' hypothesis that differential item format performance could be due to differences in motivation. Importantly, none of these prior studies investigated whether systematic differences in other item characteristics between CR and MC items (e.g., content or difficulty) 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, 2009). A cross-country 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 an analysis of PISA 2000 and 2003 mathematics data for US students, Liu and Wilson (2009) report that differences between male and female students varied depending on item format and math content assessed. However, in contrast to the other studies reporting a male advantage on MC items, they found differences were smallest (i.e., least male-favoring) for standard MC items relative to other item formats. The authors did not investigate whether item format differences were related to 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. 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. Overall, however, 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. As noted above Taylor and Lee (2012) identified some 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 assess whether these additional factors could explain 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 at least in part to female students providing longer answers with more detail, even when these details did not necessarily lead to more correct responses (Lane et al., 1996; Pomplun & Capps, 1999; Pomplun & Sundbye, 1999). Others have hypothesized that male students may be more willing 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.

To further our understanding of gender by item format differences, this study uses DIF analyses to address 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 US 15-year-olds participating in PISA reading and mathematics tests?
- 2) Is the magnitude of differences consistent with the test-level findings in prior studies?
- 3) Are gender by item format differences associated with other properties of the test items?

The first question provides a chance to directly test whether there are gender by item format differences among recent, nationally representative samples of US 15-year-olds, thus providing more up to date results relative to many of the studies reviewed above. The second question is intended to facilitate comparisons between the item-level DIF results and overall effects at the test score level. Finally, the third question is intended to help evaluate whether item format effects are due to construct relevant factors. This study contributes to the literature by combining a DIF analysis that is 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 US samples of students participating in the 2009, 2012, and 2015 administrations of PISA obtained from the OECD

PISA Database (<https://www.oecd.org/pisa/data/>). PISA is an international large-scale assessment assessing student skills in the domains of reading, mathematics, and science. The PISA assessments are intended to assess student “literacy” in these three domains, 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.

Item Format and Content

Information about item content and features were obtained from the PISA technical reports. 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. Table 2 reports the number of items overall and by format for each year and subject. Some items were used in multiple years. 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. There were 85

unique items in math with 187 item-by-year observations, and 104 unique items in reading with 233 item-by-year observations.

Table 2.

Summary of Item and Student Samples, by Year and Subject

	Math			Reading		
	2009	2012	2015	2009	2012	2015
Number of Items						
All	35	84	68	101	44	88
CR	19	51	39	53	24	46
MC	16	33	29	48	20	42
Item Means						
All	0.44	0.44	0.42	0.58	0.59	0.58
CR	0.36	0.37	0.37	0.59	0.63	0.58
MC	0.52	0.54	0.48	0.58	0.54	0.58
Standardized Mean Differences						
All	-0.16	-0.07	-0.13	0.19	0.21	0.19
CR	-0.08	-0.02	-0.09	0.20	0.24	0.21
MC	-0.20	-0.10	-0.14	0.14	0.11	0.14
Students						
N	3634	4946	2366	5230	3398	2346
% Female	49.2%	49.4%	50.5%	48.7%	49.1%	49.1%

Note. MC=multiple-choice. CR=constructed-response. Item means report the average item percent correct scores. Standardized mean differences report standardized mean difference in student percent correct scores between female and male students; positive values indicate females scored higher than males.

Next, each reading item was categorized along two dimensions based on the reading process assessed and the format of the text used in the item, while each math item was categorized along two dimensions based on the mathematics content area and mathematical process assessed.⁵ 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

⁵ Additional item features were reported in the technical documentation for both subjects but were not reported consistently across all three years and hence were not included here.

whether text accompanying the item was continuous, non-continuous, mixed or had multiple formats. The math items were coded as assessing concepts related to either “space and shape,” “change and relationships,” “quantity,” 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.”⁶ The appendix reports the count of items by format and additional item features.

Student Samples

PISA selects nationally representative samples of 15-year-olds attending educational institutions in the fall for participation. 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 US, approximately 5,000 students from approximately 165 schools participated in PISA each year, although not every student completed tests in both math and reading. 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.” Although it would be valuable to provide empirical evidence about a more complete range of gender identities, these data are not available. Table 2 reports the student sample sizes and percent of students identifying as female. Students were 15.8 years old on average and the majority (approximately 70%) were beginning 10th grade in each administration.

Descriptive Statistics

⁶ These process classifications were developed by content area experts for use with the 2012 and 2015 PISA administrations. Because 35 of the 36 items from 2009 were used in 2012, the classifications were applied to the 35 2009 items for the current analyses.

To construct the final sample within each year-subject combination, the data were limited to students with at least one non-missing response and items with at least one non-missing response. Items that were not administered to a student due to the matrix sampling were treated as missing by design. Items that a student skipped (3% of all item responses) or did not reach (0.8% of all item responses) were coded as an incorrect response. One mathematics item answered correctly by less than 1% of the total sample in 2015 was removed from the analysis. Table 2 reports the average percent correct score for each item type by year and subject for the final sample and indicates two patterns. First, the mathematics items were more difficult on average than the reading items for US 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 across test administrations.⁷ 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, on average, 0.09 SD more male-favoring when calculated using MC items rather than CR items. Although these

⁷ A percent correct score for each student was calculated as the average percent of points earned for each item. Polytomous items were worth up to two points. 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: $d = \frac{m_f - m_m}{\frac{sd_f + sd_m}{2}}$.

preliminary calculations do not fully account for the matrix sampling design, they are consistent with hypothesized gender by item format effects.

Methods

DIF analysis was used to study the presence of gender by item format effects 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. Thus uniform DIF for each item was estimated using a multi-facet Rasch item response theory (IRT) model implemented in the R package TAM (Robitzsch et al., 2022). 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. Using an IRT model to estimate DIF has two advantages in this context: adjusting for different items students were administered due to matrix sampling and producing item-specific DIF statistics that can be used to explore potential explanations for the item format effects.

Student item responses were analyzed separately for each year and subject, with the probability of a correct response to item i for student p modeled as

$$P(X_{ip} = 1 | \theta_p, \beta_i, \zeta_i, \gamma) = \frac{\exp(\theta_p + \gamma G_p - \beta_i + \zeta_i G_p)}{1 + \exp(\theta_p + \gamma G_p - \beta_i + \zeta_i G_p)}. \quad (\text{Eq. 1})$$

The variable G_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 β_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. Item 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 to 0, i.e., $\sum \zeta_i = 0$. A partial credit model parameterization was used for the small number of polytomously scored items by including a second step parameter estimate for each item but continuing to estimate a single ζ_i for the item. Model parameters were estimated via marginal maximum likelihood using the package TAM (Robitzsch et al., 2022) in the R computing environment (R Core Team, 2022).

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. In the present study, however, the goal is to identify systematic patterns of DIF across items rather than to identify individual items that may need to be revised. Thus, after fitting the model in Equation 1 for each year and subject, the $\hat{\delta}_i$ estimates for all items were analyzed in two ways to study item format effects.

To address research questions one and two, summary statistics for the $\hat{\delta}_i$ estimates were compared across item formats and a standardized item format effect was calculated. The standardized item format effect was calculated as the difference in average DIF between CR and MC items, relative to the standard deviation of θ :

$$FormatEffect = \frac{\bar{\delta}_{CR} - \bar{\delta}_{MC}}{sd(\theta)}. \quad (Eq. 2)$$

When IRT is used to scale a test, the 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. The format effect in Equation 2 indicates how much we would expect the standardized mean difference between female and male students to change if moving from a test with entirely MC items to one with entirely CR items. A positive format effect indicates that moving from a test with entirely MC items to entirely CR items would favor female students.

To address research question three, least squares multiple linear regression models were used to examine whether differences in DIF between MC and CR items might be due to other relevant factors that differ between item types. The following regression model was estimated separately for each subject:

$$\hat{\delta}_{iy} = \alpha_0 + \alpha_1(isCR_i) + \omega \mathbf{X}_{iy} + \eta_y + \epsilon_{iy}. \quad (Eq. 3)$$

The $\hat{\delta}_{iy}$ is the estimated DIF coefficient of item i in year y , $isCR_i$ is an indicator equal to 1 for CR items and 0 for MC items, \mathbf{X}_{iy} is a vector of item covariates, and η_y are year fixed-effects. Item covariates include estimated item difficulty 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 (i.e., the difference in average DIF between CR and MC items), conditional on other item covariates included in the model. Models were estimated with and without \mathbf{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. All data files and computer code to reproduce the analyses in this paper are available from the authors upon request.

Results

Table 3 summarizes the average and range of DIF estimates by year, subject, and item format. The average DIF values by format range from -0.09 to 0.07 across administrations. Consistent with the hypothesized format effects, the average DIF for CR items was positive in every administration ($M = 0.05$ logits overall) and average DIF for MC items was negative in every administration ($M = -0.06$ logits overall). Although the average DIF showed a consistent pattern, there was substantial variation in item specific 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. Although this provides some preliminary evidence that item format effects may differ for computerized versus paper tests, we should be cautious about generalizing from a single year of data and further investigation is warranted.

The standardized format effects in Table 3 range from 0.02 (2015 math) to 0.13 (2012 reading), with an average of 0.08 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 by 0.08 SD. For reference, the estimated SD of θ in Equation 1 ranged from 1.15 to 1.28 (Mean=1.21) across administrations, and the standardized mean differences in overall theta scores between female and male students ranged from -0.07 to -0.18 SD in math and 0.20 to 0.26 SD in reading. A change in the standardized mean difference of 0.08 SD could have a practically meaningful impact on inferences about relative performance of female and

male students. In an operational testing context where the focus is on identifying biased test items, a single item with DIF of 0.05 to 0.10 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 3.

DIF Summary Statistics and Format Effects, by Year and Subject

Subject	Year	CR			MC			Format Effect	d
		Mean	Min.	Max.	Mean	Min.	Max.		
Math									
	2009	0.06	-0.37	0.51	-0.07	-0.57	0.35	0.10	-0.18
	2012	0.06	-0.57	0.86	-0.09	-0.61	0.50	0.12	-0.07
	2015	0.01	-0.46	0.62	-0.01	-0.46	0.47	0.02	-0.16
Reading									
	2009	0.05	-0.58	0.44	-0.05	-0.54	0.31	0.08	0.20
	2012	0.07	-0.54	0.37	-0.08	-0.43	0.54	0.13	0.26
	2015	0.03	-0.78	0.61	-0.04	-0.58	0.95	0.05	0.23

Note. CR=constructed-response; MC=multiple-choice; *d*=standardized mean difference of theta estimates; positive values indicate females scored higher than males. Standard deviations of theta ranged from 1.15 to 1.28 (average 1.21) across test administrations. Positive DIF values indicate items easier for female students.

Table 4 presents the regression analysis results. In both subjects the item format differences remained after adjusting for other item covariates. The set of item covariates explained between 13-16% of the variance in item DIF across subjects. Because the item DIF estimates are not independent, the standard errors and p-values should be interpreted cautiously but are provided as reference values. In math, the conditional format difference was slightly larger (0.13 versus 0.09) when controlling for difficulty, math content, and math process

assessed and would be statistically significant at the $p < 0.01$ level. In reading, the conditional format difference was slightly smaller (0.08 versus 0.10) when controlling for difficulty, reading process assessed, and text format, and would be statistically significant at the $p < 0.05$ 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 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).

Table 4.*Regression Model Estimates Predicting Item DIF*

<i>Predictors</i>	Math 1			Math 2			Reading 1			Reading 2		
	<i>Est.</i>	<i>SE</i>	<i>p</i>	<i>Est.</i>	<i>SE</i>	<i>p</i>	<i>Est.</i>	<i>SE</i>	<i>p</i>	<i>Est.</i>	<i>SE</i>	<i>p</i>
Intercept	-0.06	0.05	0.19	-0.02	0.06	0.73	-0.05	0.03	0.11	-0.04	0.05	0.46
Is CR	0.09	0.04	0.02	0.13	0.04	<0.01	0.10	0.04	0.01	0.08	0.04	0.02
Difficulty				-0.02	0.01	0.11				-0.06	0.01	<0.01
Quantity				0.00	0.05	0.93						
Space and Shape				-0.07	0.05	0.17						
Uncertainty and Data				-0.02	0.05	0.72						
Formulate				-0.11	0.05	0.01						
Interpret				-0.03	0.05	0.46						
Integrate and Interpret										0.01	0.05	0.79
Reflect and Evaluate										0.05	0.05	0.30
Text Format Mixed										-0.11	0.06	0.07
Text Format Multiple										0.10	0.12	0.38
Text Format Non-continuous										-0.06	0.04	0.17
Observations	186			186			233			233		
R ²	0.03			0.13			0.03			0.16		

Note. 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.

Discussion

The results of this study provide consistent evidence that among recent samples of US high school students, 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. The magnitude of the item format differences was similar across subjects, although slightly smaller for the computerized tests used in 2015, a difference that suggests further investigation would be warranted. Although average format differences were consistent, there was substantial variability in the relative differences across individual items; that is, although the difference was observed

consistently across years and subjects, male students do not always do better on all MC items and vice versa for CR items.

Regarding the second research question, the observed format differences in the present analysis were smaller than the effects in the test-level analyses presented and summarized above. The average item format effect in the present analysis was 0.08 SD, indicating that the standardized mean difference in scores would be 0.08 SD more male-favoring on a test with entirely MC items relative to one with entirely CR items. The item format effect estimates in the analyses by Reardon et al. (2018) were approximately 0.20 SD on average across grades and subjects, and potentially larger in the examples presented in Table 1. However, format effects of 0.08 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 compared in prior analyses measure different knowledge and skills in addition to using different item formats.

Regarding the third research question, 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 across administrations 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, we should be cautious about generalizing the results here to the broader population of high school students and to other tests beyond the low-stakes PISA tests. The use of the IRT DIF model assumes 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 other aspects of students' identity or experiences such as their racial/ethnic identity, family economic situation, or school environment were associated with differential performance. The study also did not investigate whether these differences exist in other countries. Investigating potential item format differences across groups defined by other identities, or between intersections of these identities, 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. Here I briefly describe three implications. 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 potential 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 male students were more likely to omit item responses across all years, subjects, and item formats, and this difference was more pronounced for CR than for MC items. 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 coding of omitted responses. As a sensitivity analysis, the item format effects were re-estimated after replacing omitted responses (including not reached items) with the modal non-missing

response for each item. The substantive conclusions and direction of DIF across item formats remained the same, although the average standardized format effect was slightly smaller.

Second, regardless of the cause of the 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.

Third, a challenge highlighted by the analyses here is 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 differences. Part of the PISA development and validation process includes monitoring gender DIF, as do the development and validation of the state tests used in the test level analyses cited above. 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.

References

- 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>
- 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.
- Holland, P. W., & Wainer, H. (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. (2009). 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>
- 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. (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*, 1–11.

<https://doi.org/10.3102/0013189X18762105>

Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test Analysis Modules* (R package version 4.0-16). <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>

- Shear, B. R., & Reardon, S. F. (2019). *HETOP: Stata module for estimating heteroskedastic ordered probit models with ordered frequency data* (Version 3). Statistical Software Components. <https://ideas.repec.org/c/boc/bocode/s458287.html>
- 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/>

Appendix

Table A1.*Number of Items by Item Properties, Year, and Format*

Subject	Category	2009		2012		2015	
		CR	MC	CR	MC	CR	MC
Math	All Items	19	16	51	33	39	29
	Change and Relationships	6	3	16	5	11	5
	Quantity	6	5	11	10	8	10
	Space and Shape	5	3	14	7	10	6
	Uncertainty and data	2	5	10	11	10	8
	Employ	10	4	22	14	16	13
	Formulate	6	4	20	7	15	5
	Interpret	2	8	9	12	8	11
	Reading	All Items	53	48	24	20	46
Access and retrieve		16	7	7	3	14	8
Integrate and interpret		17	36	9	15	15	31
Reflect and evaluate		20	5	8	2	17	3
Continuous		31	31	13	13	27	27
Mixed		4	7	2	2	2	5
Multiple		1	0	1	0	1	2
Non-continuous		17	10	8	5	16	8

Note. CR=constructed response; MC=multiple-choice.