The Relationship Between Test Item Format and Gender Achievement Gaps on Math and ELA Tests in Fourth and Eighth Grades

Sean F. Reardon[1], Demetra Kalogrides[1], Erin M. Fahle[1], Anne Podolsky[2], and Rosalía C. Zárate[1]

Prior research suggests that males outperform females, on average, on multiple-choice items compared to their relative performance on constructed-response items. This paper characterizes the extent to which gender achievement gaps on state accountability tests across the United States are associated with those tests' item formats. Using roughly 8 million fourth- and eighth-grade students' scores on state assessments, we estimate state- and district-level math and reading male-female achievement gaps. We find that the estimated gaps are strongly associated with the proportions of the test scores based on multiple-choice and constructed-response questions on state accountability tests, even when controlling for gender achievement gaps as measured by the National Assessment of Educational Progress (NAEP) or Northwest Evaluation Association (NWEA) Measures of Academic Progress (MAP) assessments, which have the same item format across states. We find that test item format explains approximately 25% of the variation in gender achievement gaps among states.

Keywords: achievement gaps; assessment; correlational analysis; English language arts; gender studies; mathematics; quasi-experimental analysis; test item format; test theory/development; testing

Studies of gender achievement gaps in the United States show that on average, females outperform males on reading/English language arts (ELA) tests and males outperform females on math tests (Chatterji, 2006; Fryer & Levitt, 2009; Husain & Millimet, 2009; Lee, Moon, & Hegar, 2011; Penner & Paret, 2008; Robinson & Lubienski, 2011; Sohn, 2012). These test-based gender achievement gaps are often used to help understand how gender norms and stereotypes shape students' lives and shed light on gender disparities in educational opportunity. But what if the conclusions we draw are sensitive to how we measure gender achievement gaps on standardized tests?

Gender achievement gaps are typically estimated by comparing male and female students' average total scores on an assessment. If a test measures a unidimensional construct, so that gender gaps do not vary on different items or parts of the test, this approach is appropriate. If, however, gender differences in achievement vary among the set of skills tested, then gender gaps computed from the overall scores will depend on the mix of skills measured by the test.

Prior research suggests that we should be concerned about the latter. There is evidence of a relationship between gender achievement gaps and item format—gaps are often more male-favoring on tests with more multiple-choice items and more female-favoring on tests with more constructed-response items. This pattern may be due to gender differences on various construct-relevant skills—the skills intended to be measured by the test—and the use of different item types to assess the different skills. Alternatively, the pattern may be due to gender differences in the ancillary, construct-irrelevant skills required by the different item types (e.g., the handwriting skills required for essay questions). Either way, a relationship between test item format and gender achievement

[1]Stanford University, Stanford, CA
[2]Learning Policy Institute, Palo Alto, CA

Educational Researcher, Vol. 47 No. 5, pp. 284–294
DOI: 10.3102/0013189X18762105

gaps suggests that a single summative gap measure may lead to inaccurate assessments of the magnitude of gender achievement gaps, inefficiencies in the efforts to close them, and distorted comparisons of gender achievement gaps across state tests that weight the dimensions differently in overall scores.

In this paper, we build on existing work by systematically characterizing the relationship between test item format and estimated gender achievement gaps in performance. We use the scores on state accountability assessments of roughly 8 million students tested in fourth and eighth grades in ELA and math during the 2008–2009 school year to estimate state- and district-level subject-specific gender achievement gaps on each state's accountability tests. We then show that these measured gaps are strongly associated with the proportion of the total score that is derived from multiple-choice versus constructed-response items. This relationship holds even when we control for each state or district's gender gap estimated using a separate test that is the same across all states and districts. Although we cannot determine whether the observed variation in the gap is due to gender differences in construct-relevant or -irrelevant skills associated with item format, our analysis shows that format explains approximately 25% of the variation in state- and district-level gender achievement gaps in the United States.

Background

We often think of achievement tests as unidimensional, which leads to the conclusion that a single measure adequately captures gaps in performance between student subgroups on a test. However, achievement tests are often complex and measure multiple related dimensions of a broad construct. Consider a state ELA assessment. The assessment may measure vocabulary, writing, and reading comprehension—correlated but disparate dimensions of ELA skills. For a single achievement gap to sufficiently characterize differences in performance, the achievement gaps on the different dimensions of the assessment (e.g., on the vocabulary items, writing items, and reading comprehension items) must be the same. If the gaps are not the same, however, then the weighting of the dimensions in the total score will impact the size of the overall achievement gap.

But is the assumption that the gender performance gaps are constant across all dimensions of an assessment reasonable? Prior empirical research suggests not. It shows that gender achievement gaps can be sensitive to item format, where item format is defined by the mode(s) of response an item requires. These studies focus on the difference in performance on multiple-choice items—items that require students to select a response from a list of possible answers—versus constructed-response items—items that require students to write their own answer (ranging in length from a sentence to an essay) in response to a prompt (National Center for Education Statistics [NCES], 2009a, 2009b). This research generally shows that male students score higher, on average, than female students on the multiple-choice portions of tests, whereas female students score higher, on average, on the written portions of tests (Beller & Gafni, 2000; Bolger & Kellaghan, 1990; DeMars, 1998, 2000; Gamer & Engelhard, 1999; Hastedt & Sibberns, 2005; Hyde, Fennema, & Lamon, 1990; Lafontaine & Monseur, 2009; Lindberg, Hyde, Petersen, & Linn, 2010; Mullis, Martin, Fierros, Goldberg, & Stemler, 2000; Routitsky & Turner, 2003; Schwabe, McElvany, & Trendtel, 2015; Taylor & Lee, 2012; Willingham & Cole, 2013; Zhang & Manon, 2000).

In a meta-analysis of math assessments, Lindberg et al. (2010) found that on average, the male-female achievement gap on multiple-choice math items was .18 standard deviations larger than the corresponding gap on short-response items and .22 standard deviations larger than that on extended-response items. Taylor and Lee (2012) found that multiple-choice questions generally favor males and constructed-response questions generally favor females for Grades 4, 7, and 10 on the Washington state reading and math tests. Moreover, Schwabe et al. (2015) found that among 10- and 15-year-old students who participated in two large-scale reading assessments (the German PIRLS in 2011 and the PISA in 2009), females scored higher than males, on average, on constructed-response reading items relative to the difference in their scores on other items. This evidence is not conclusive, however; some earlier studies have found inconsistent results using different assessments (Beller & Gafni, 2000) or no gender differences (Dimitrov, 1999; O'Neil & Brown, 1998; Roe & Taube, 2003; Routitsky & Turner, 2003).

There are two likely explanations for why gender achievement gaps may vary with item format. First, an assessment may use different item formats to measure different construct-relevant dimensions of skills, and gender achievement gaps may vary across those dimensions (e.g., Taylor & Lee, 2012). For example, if an ELA assessment measures writing skills using constructed-response items and vocabulary skills using multiple-choice items, then a more male-favoring gap on multiple-choice items could be an artifact of males having better average vocabulary skills relative to their writing skills compared to females. Taylor and Lee (2012) found patterns consistent with this explanation in an analysis of the content of the multiple-choice and constructed-response items on which they observed gender differences in performance. In reading, they found that males tended to perform better on items that ask students to identify reasonable interpretations and analyses of informational text. On average, females performed better on items where they were asked to make their own interpretations and analyses of literary and informational text, supported by text-based evidence. Geometry, probability, and algebra items favored males, on average, while statistical interpretation, multistep problem solving, and mathematical reasoning items generally favored females.

A second explanation is that the relationship may be driven by different gender gaps in the ancillary, construct-irrelevant skills required to answer multiple-choice versus constructed-response items.[1] Abedi and Lord (2001) and Abedi, Lord, and Plummer (1997), for example, note that language comprehension skills affect student performance on math tests. They argue, however, that reading comprehension should be understood as construct-irrelevant in tests designed to assess math skills per se. Other potential ancillary skills may include guessing for multiple-choice items or handwriting for constructed-response items. In this case, we might interpret males' lower average performance on constructed-response items relative to their performance on multiple-choice items as resulting from their poorer handwriting on constructed-response items or higher propensity to guess on multiple-choice items. Prior literature finds little support for this

hypothesis, although few potential ancillary skills have been explored. Ben-Shakhar and Sinai (1991) and von Schrader and Ansley (2006) hypothesize that on average males and females may perform differently on multiple-choice questions because males are more likely to guess whereas females have higher omission rates. However, neither study found that guessing or omission explained the gaps on the assessments.

Such research suggests that we should be concerned about a relationship between item format and gender achievement gaps, particularly in the context of high-stakes state standardized assessments. These tests are used by many school districts to assign students to courses, and this correlation may have meaningful consequences for students. However, only a few prior empirical studies on gender differences in performance by item format have analyzed recent state accountability tests, and those that do generally focus on a single state (Dimitrov, 1999; Gamer & Engelhard, 1999; O'Neil & Brown, 1998; Taylor & Lee, 2012). The goal of our paper is to quantify the extent to which the proportion of multiple-choice items is related to male-female gaps on current state assessments in order to understand whether a single summative gap measure masks important gender differences within these assessments.

Research Aims and Hypotheses

We seek to answer two primary research questions:

Research Question 1: Is there a systematic relationship between the format of test questions and differences in males' and females' test scores on state accountability tests?

Research Question 2: Does the association vary across grades and subjects?

To answer the first question, we model the gender achievement gap in test scores on mandatory state ELA and math assessments in fourth and eighth grades as a function of the proportion of the total score that is based on constructed-response items. We hypothesize that the proportion of the score from multiple-choice items will be associated with more positive (male-favoring) gender achievement gaps and constructed-response items will be associated with more negative (female-favoring) gaps, as suggested by prior research.

To answer our second question, we formally test whether the relationship varies significantly across grades or subjects. Based on the prior literature, we hypothesize that there will be a significant relationship in each of the grades and subjects included in our analysis. However, it is unclear that the relationship will be the same across grades and subjects. There may be meaningful differences in the structure of constructed-response questions (e.g., single word response vs. essay questions) or multiple-choice questions (e.g., number of response options) that are used across grades or subjects. For example, constructed-response questions in fourth grade may require shorter responses and therefore less handwriting or other ancillary skills than the constructed-response questions in eighth grade. For multiple-choice items, the number of response options may be larger in eighth grade than fourth grade and therefore may attenuate the benefits of ancillary skills like guessing in eighth grade compared to fourth grade. Alternatively, the content assessed by items of different types may vary between grades or subjects. For example, multiple-choice items may be used to assess arithmetic in fourth grade but algebra in eighth grade, and the gender gaps may be smaller on algebra than arithmetic. Although our analyses cannot test these specific hypotheses, because we do not have access to item-level data for state assessments, we are able to test whether the association between item format and gender gaps varies across grades and subjects.

Data

We use student achievement data from three primary sources: NCES ED*Facts* Database, National Assessment of Educational Progress (NAEP) data, and Northwest Evaluation Association (NWEA) Measures of Academic Progress (MAP) assessment data.

ED*Facts* is a U.S. Department of Education initiative to centralize performance data from K–12 state education agencies. The NCES provided us the ED*Facts* data via a restricted use data license. The data consist of categorical proficiency data (e.g., percentages of students scoring "Below Basic," "Basic," "Proficient," and "Advanced") for each state and school district, disaggregated by gender, grade, subject, and year. We use data from 47 states[2] in Grades 4 and 8 in the 2008–2009 school year given that we only have state tests' item format information for that year and those grades.

The information on each state's test item format comes from a series of NAEP reports (one report for each state) (NCES, 2011). The reports provide information on each state's accountability tests for reading and math assessments in 2008–2009 for Grades 4 and 8. Specifically, the reports indicate the proportion of the total score that is based on items of each of several mutually exclusive formats: multiple choice, short constructed response, extended response, performance tasks, or other. We use these proportions as our key explanatory variables.[3]

Table 1 summarizes the average proportion of the test score based on items of each format type on state math and ELA assessments in 2008–2009 in Grades 4 and 8. Multiple-choice questions comprise, on average, approximately 80% of the proportion of the total score on the state assessments, ranging from approximately 39% to 100% of the proportion of score across states.

The NAEP data set includes student achievement in math and reading assessments of representative samples of fourth- and eighth-grade public school students in each state. The NAEP assessments have a common format and common content across all states. The data include roughly 3,000 to 4,000 student test scores for each state-grade-subject cell.[4]

The NWEA test database includes math and reading test scores for the majority of students in about 10% of all districts nationwide. The NWEA MAP assessments are computer-adaptive multiple-choice tests. Although the specific items included in students' tests differ among students, item response theory scoring of the tests yields scores on a common metric for all students in the country who take the test. Districts administering the NWEA tests typically assess all students in the district in a given grade; we exclude a small number of districts where fewer

Table 1
Means and Standard Deviations of Test Item Properties

	Math				English Language Arts			
	Grade 4		Grade 8		Grade 4		Grade 8	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Proportion of score								
Multiple choice	.818	.185	.787	.200	.821	.178	.818	.179
Short response	.087	.113	.088	.112	.098	.141	.102	.144
Extended response	.091	.131	.096	.125	.080	.132	.080	.132
Performance tasks	.002	.008	.002	.009	.000	.000	.000	.000
Other	.002	.009	.028	.073	.000	.000	.000	.000
Proportion of items								
Multiple choice	.903	.109	.876	.144	.922	.081	.919	.085
Short response	.059	.074	.065	.084	.052	.071	.054	.072
Extended response	.035	.052	.039	.054	.023	.035	.025	.038
Performance tasks	.001	.004	.001	.004	.000	.000	.000	.000
Other	.002	.011	.019	.054	.002	.017	.002	.015

than 90% of students have valid NWEA test scores. The final sample of districts for which we have NWEA test data contains 794 and 665 district observations for ELA Grades 4 and 8, respectively, and 777 and 696 district observations for math in Grades 4 and 8, representing approximately 7% of students in Grades 4 and 8 nationwide in 2009.

Methods

We estimate male-female achievement gaps in each state or district using the *V*-statistic (Ho, 2009; Ho & Haertel, 2006; Ho & Reardon, 2012). As Ho and colleagues explain, *V* is akin to Cohen's *d*, the difference in means between two groups divided by their pooled standard deviation. The distinction between *V* and *d* is that *V* depends only on the ordered nature of test scores (it does not assume scores represent an interval scale, as *d* does) and can be computed accurately from highly coarsened data (Ho & Reardon, 2012; Reardon & Ho, 2015). The *V*-statistic is a measure of the degree of nonoverlap between two distributions; it is insensitive to how achievement is scaled and so can be used to compare gaps on tests that measure achievement in different metrics. These features of *V* are useful for our analyses since each state fields a different accountability test and reports scores in different scales.

We estimate state and district male-female achievement gaps (*V*) from the ED*Facts*, NAEP, and NWEA data using the methods described by Ho and Reardon (2012; Reardon & Ho, 2015).[5] A positive gap indicates that males outperform females on average in a given state or district; a negative gap indicates that females outperform males.

Models

To understand the relationship between male-female achievement gaps and item format, we begin with a simple model. If males and females perform differentially well on multiple-choice and constructed-response test items, then the measured gap will depend in part on the proportion of score based on items of each type on the test.[6] We can write the achievement gap (G) as measured by the state accountability test t as:

$$\begin{aligned} G_{st} &= \gamma_s + \delta\, p_{st} + u_{st} \\ &= \gamma + \delta\, p_{st} + \left(v_s + u_{st}\right) \\ &= \gamma + \delta\, p_{st} + u^*_{st}, \end{aligned} \tag{1}$$

where p_{st} is the proportion of score based on non–multiple choice items on test t in state/district s; $\gamma_s = \gamma + v_s$ is the male-female gap in achievement in state/district s if measured by a test that is common across states and that contains only multiple-choice items (γ is the average male-female gap, and v_s is the difference between the gap in state/district s and the average across states/districts); and u_{st} represents any non–test format gender bias in test t in state/district s. That is, if the average errors in the test's measurement of males' and females' achievement are unequal for reasons unrelated to the test's item format, then u_{st} will be non-zero. For example, if the test contained items whose content were culturally biased toward females, then u_{st} might be negative. In Model 1, δ is the parameter of interest; it describes the association between the measured gender gap and the item format of a given test. We wish to test the null hypothesis that $\delta = 0$, that is, that item format does not affect measured gender gaps.

Our estimate of δ from Model 1 will be biased if the proportion of non–multiple choice items on a state's test is correlated with the size of the gender gap measured by a common test (i.e., if $p_{st} \sim v_s$) or if other sources of error in the measured achievement gaps are correlated with the proportion of non–multiple choice items on the test (if $p_{st} \sim u_{st}$). We improve on Model 1 and reduce the first source of bias in our estimates by using gender achievement gaps among students in state or district s measured by two tests of the same subject and different item formats that are administered to the same population of students.[7] One test

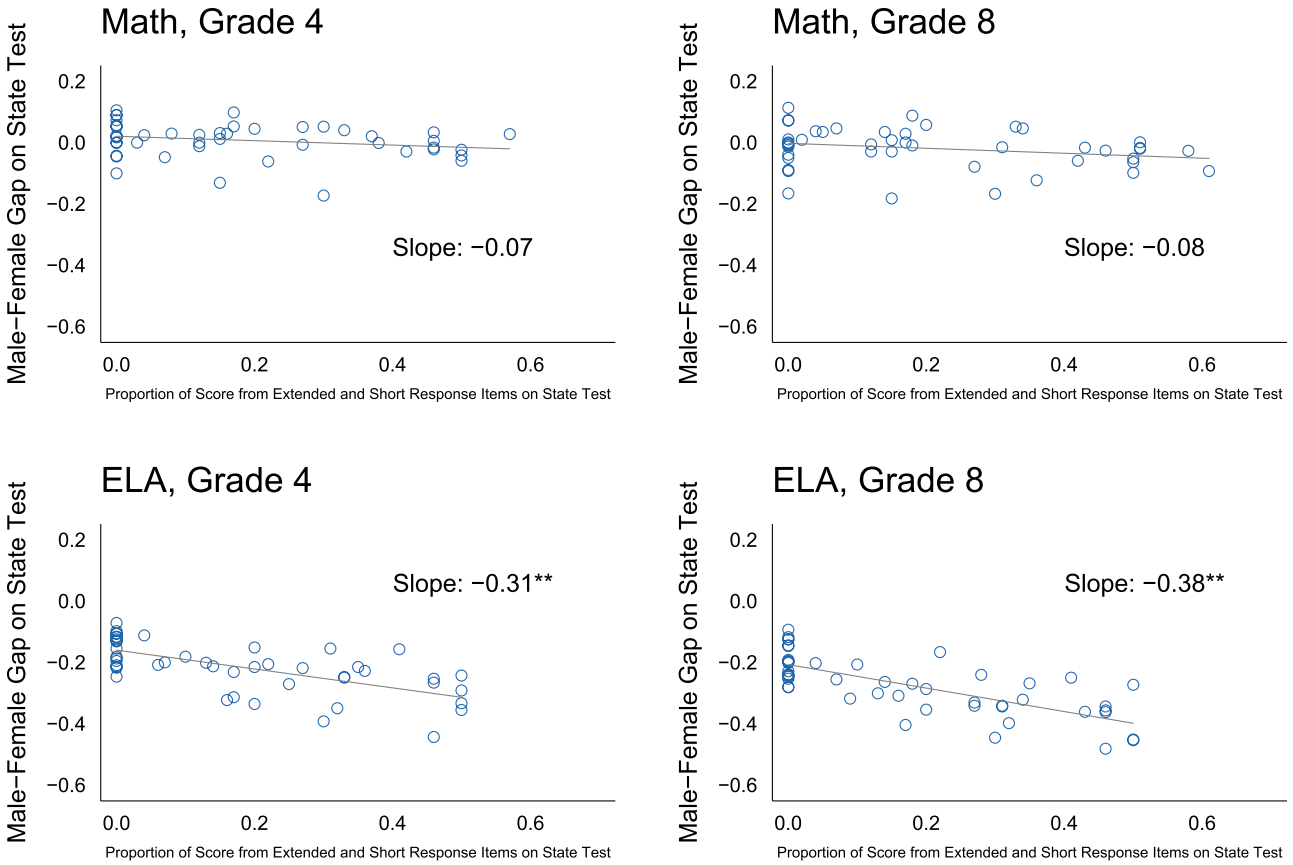

FIGURE 1. *Male-female achievement gaps versus proportion of score from constructed-response items on state test, by grade and subject.*

(denoted $t = a$) is a state accountability test (from ED*Facts*, as used previously) whose format varies among states. The other is a national test (denoted $t = n$) that is identical in each state/district. This is either the NAEP or NWEA MAP assessment, depending on if the analysis is at the state or district level. Note that this implies $p_{sn} = p_n$ is a constant since test n is identical in each state/district. Let T_{st} be an indicator variable for the state accountability test (so $T_{sa} = 1$ and $T_{sn} = 0$) and define $\alpha = -\delta p_n$. Then we can express the gap on test t in state or district S as:

$$\begin{aligned} G_{st} &= \gamma_s + \delta p_{st} + u_{st} \\ &= \gamma_s + (\delta p_{sa}) T_{st} + (\delta p_n)(1 - T_{st}) + u_{st} \\ &= [-\alpha + \gamma_s] + \alpha T_{st} + \delta (p_{sa} \cdot T_{st}) + u_{st} \\ &= \gamma_s^* + \alpha T_{st} + \delta (p_{sa} \cdot T_{st}) + u_{st}. \end{aligned} \quad (2)$$

We can estimate δ by fitting Model 2 using state or district fixed effects if we know p_{sa}, the proportion of score from non–multiple choice items on each state/district accountability test. We do not need to know p_n because it is a constant and so is absorbed in α. The use of state or district fixed effects in the models means that we are essentially controlling for the gender gap in each state as measured by a common test in the models.[8] Under the assumption that $u_{sa} \perp p_{sa}$—that other sources of gender bias in the accountability tests are not correlated with the proportion of score from multiple-choice items on those tests—we can estimate δ without bias. If we find that $\delta \neq 0$, this indicates that there are systematic gender differences in students' performance on items of different types.

We fit four versions of Model 2 at the state and district levels: separately for each grade and subject combination and then pooling across grades within subjects, across subjects within grades, and across all four grade-subject combinations. In these latter models, we include state or district by subject, state or district by grade, or state or district by grade by subject fixed effects as appropriate, and we allow α to vary across grades and subjects (because p_n may vary across grades and subjects). In all models, we weight each observation by the inverse of the sampling variance of the gap estimate. For the district-level models, we cluster the standard errors in these models at the state level because there are multiple observations per state.

Results

Figure 1 plots the estimated average male-female achievement gap on each state's accountability assessment against the proportion of the score that is derived from constructed-response items (both short and extended response). In all subject and grade combinations, there is a negative relationship between the proportion of constructed-response items and the male-female gap indicating that gaps are more female-favoring on tests with higher proportions of constructed-response questions. This is consistent with our previous hypothesis. The slopes of the regression lines in Figure 1, corresponding to δ in Equation 1, indicate that the proportion of constructed-response items is more strongly associated with the male-female achievement gap in ELA than it is in math; in eighth-grade math, there does not

Table 2
Relationship Between Proportion of Score From Multiple-Choice Items on State Tests and the Size of Male-Female Gaps, State-Level Analyses

	Pooled Across Grades and Subjects		Pooled Across Subjects		Pooled Across Grades		Math		English Language Arts	
	Without Audit Test	With Audit Test	Grade 4	Grade 8	Math	English Language Arts	Grade 4	Grade 8	Grade 4	Grade 8
Model 1										
Proportion short response (SR) + extended response (ER)	–.214***	–.202***	–.174**	–.236***	–.116**	–.289***	–.127**	–.104	–.221*	–.365**
	(.027)	(.051)	(.053)	(.066)	(.036)	(.080)	(.042)	(.069)	(.085)	(.107)
p value from test that coefficients are equal across grades/subjects	.000	.096	.283	.046	.784	.168				
Residual variance explained by test items		.235	.236	.244	.175	.310	.259	.115	.252	.376
Model 2										
Proportion SR	–.274***	–.161*	–.143^{+}	–.184^{+}	–.081	–.236*	–.065	–.099	–.204^{+}	–.274*
	(.080)	(.076)	(.085)	(.092)	(.058)	(.102)	(.064)	(.123)	(.120)	(.119)
Proportion ER	–.173*	–.238***	–.200***	–.284***	–.143*	–.343***	–.174***	–.107	–.237*	–.465***
	(.077)	(.052)	(.054)	(.080)	(.061)	(.084)	(.042)	(.127)	(.097)	(.114)
p value from test that SR = ER	.355	.390	.560	.422	.535	.319	.171	.971	.803	.150
Residual variance explained by item format		.244	.242	.255	.183	.323	.291	.115	.253	.409
N	188	376	190	186	184	192	94	90	96	96
Fixed effects included in model	Grade-subject	State-grade-subject	State-subject		State-grade		State	State	State	State
Interaction terms included in model		Test-grade-subject	Test-subject		Test-grade		—	—	—	—

Note. All models are weighted by the inverse of the sampling variance of the gender gap. Standard errors (in parentheses) are clustered by state. The models include data from 2008–2009 ED*Facts* and National Assessment of Educational Progress (NAEP) data sources from Grades 4 and 8. The models are restricted to state by grade cells with gap data from both ED*Facts* and NAEP. Model 1 and Model 2 are identical except that Model 1 adds the proportion of short and extended response items together while model 2 estimates coefficients for these measures separately. Both models also include the proportion of "other" (not shown) items.

appear to be a strong relationship between the two. However, even among states with tests that have the same proportion of constructed-response items, there is significant variation in the size of the gaps that is not explained by test item format. In the simple model, pooling the data across grades and subjects (shown in Column 1 of Table 2, top panel), the association between the proportion of constructed-response items and the male-female gap is –.214 (*SE* = .027, $p < .001$).

Although Figure 1 suggests that there is a relationship between test format and gender achievement gaps, at least in ELA, this simple correlation may be confounded by an unobserved factor that is correlated with both the gender achievement gap and the proportion of constructed-response items on a test. To reduce such potential bias, we fit the four variants of Model 2; these models control for gender achievement gaps measured by a common test across states and so reduce bias in our estimates that is due to between-state differences in gender gaps as measured by a common test.

Table 2 shows the results of our state-level version of this analysis using both Ed*Facts* and NAEP. The top panel reports estimates from models that combine the proportion of score based on all constructed-response questions; the bottom panel reports estimates from models with the proportions of short and extended response formats included separately. Column 1 reports estimates of the simple model, pooling observations across grades and years, that does not control for a common measure of achievement gaps (Equation 1). Column 2 reports estimates from our model controlling for the achievement gap on the NAEP test and pooling observations across grades and years (Equation 2). The estimate of δ here is –.202 *SD*. This implies that the male-female achievement gap is approximately .10 *SD* larger (in favor of males), on average, on tests that are 100% multiple-choice than tests with 50% of their score based on constructed-response questions (roughly the largest proportion on any of the state tests).

Columns 3 through 10 of Table 2 report variants of the model estimated separately by grade, subject, and subject-grade combination. These models suggest that δ is larger in ELA than math and larger in Grade 8 than 4. Column 4 suggests that there may be a stronger advantage for females on constructed-response

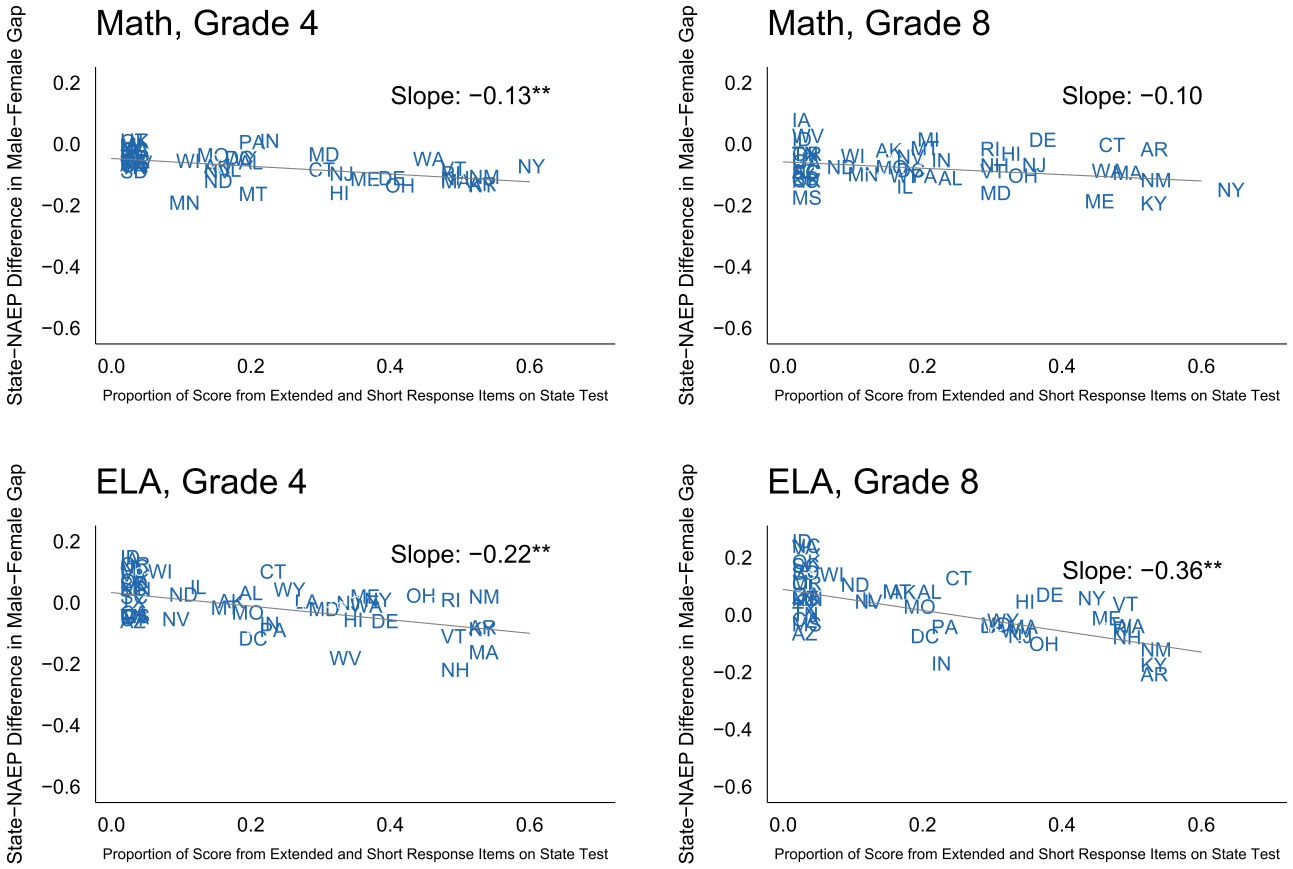

FIGURE 2. *Difference in gender achievement gap between state and National Assessment of Educational Progress (NAEP) tests by proportion score from extended and short response items on state test.*

items in ELA versus math (the p value for the test that the association is the same in math and ELA is .046, and the coefficient is smaller for Grade 8 math than ELA). However, we cannot reject the null hypothesis that δ is equal across the four grade-subject combinations (see Column 2: p = .096).

The bottom panel of Table 2 reports the results of similar models but with δ allowed to differ for short and extended response item formats. In each column, the third row reports the results of the test that the two coefficients are equal. Although the coefficient on extended response item format is larger in every model than the coefficient on short response items, the difference is never statistically significant. The most parsimonious model—Column 2 of the top panel—therefore appears to be the best fitting model. Note, however, that test format explains about 25% of the residual variance in gender achievement gaps across states after controlling for the gaps on NAEP. This suggests that there are other test-related factors (e.g., item difficulty) that may generate variation in gender gaps on state tests.

Figure 2 illustrates the relationship between the difference in the state and NAEP male-female achievement gaps and the proportion of score based on constructed-response items on the state assessment. The x-axis is the proportion of a state's test score that is based on constructed-response items. This ranges from 0, indicating the state assessment is all multiple choice, to approximately .6, indicating that 60% of the state assessment score is based on constructed-response format items. On the y-axis, a negative (positive) difference indicates that the state test gender gap is relatively more female-favoring (male-favoring) than the corresponding NAEP assessment gap. The fitted lines correspond to the associations estimated from the regression models in the top panel of Columns 7 through 10 in Table 2. In each grade and subject, the difference between the gender achievement gaps measured on the state and NAEP assessments is more negative (indicating that the state tests are relatively more female-favoring than the NAEP tests) in states where the state test scores are based more heavily on constructed-response items.

Table 3 shows the results of the analogous district-level analysis using the NWEA assessment data in place of the NAEP data. The results are very similar to those in Table 2. The estimates of δ are larger in ELA than math and larger for extended-response than short-answer items. Nonetheless, just as in the state-level analyses in Table 2, we cannot reject the null hypotheses that δ is equal across grades and subjects and for short and extended response items. The best fitting model is again the most parsimonious (Column 2, top panel). The estimated value of δ here is –.224 *SD*, roughly the same size as in Table 2, implying that gender gaps differ by .11 *SD*, on average, on tests with 0% and 50% of their score based on constructed-response items.

In sum, the models show a significant relationship between test item format and the magnitude of the male-female achievement gap. This pattern holds across state- and district-level comparisons with the use of different audit tests (NAEP and NWEA MAP). Although the estimated association appears larger in ELA

Table 3
Relationship Between Proportion of Score From Multiple-Choice Items on State Tests and the Size of Male-Female Gaps, District-Level Analyses

	Pooled Across Grades and Subjects		Pooled Across Subjects		Pooled Across Grades		Math		English Language Arts	
	Without Audit Test	With Audit Test	Grade 4	Grade 8	Math	English Language Arts	Grade 4	Grade 8	Grade 4	Grade 8
Model 1										
Proportion short response (SR) + extended response (ER)	−.302***	−.224***	−.243**	−.213*	−.090	−.351***	−.115	−.075	−.373**	−.329
	(.025)	(.059)	(.079)	(.079)	(.094)	(.080)	(.116)	(.106)	(.113)	(.093)
p value from test that coefficients are equal across grades/subjects	.000	.185	.151	.032	.723	.739				
Residual variance explained by test items		.032	.038	.028	.005	.078	.010	.003	.086	.071
Model 2										
Proportion SR	−.280***	−.213+	−.193	−.262*	−.125	−.304*	−.091	−.248	−.302+	−.313*
	(.041)	(.125)	(.139)	(.119)	(.151)	(.122)	(.141)	(.186)	(.159)	(.139)
Proportion ER	−.250***	−.236**	−.297***	−.169	−.059	−.398***	−.141	.061	−.448***	−.346**
	(.041)	(.072)	(.054)	(.129)	(.131)	(.085)	(.137)	(.163)	(.106)	(.125)
p value from test that SR = ER	.590	.890	.535	.628	.752	.530	.758	.283	.389	.864
Residual variance explained by item format		.032	.040	.029	.006	.080	.011	.010	.090	.071
N	2,932	5,864	3,142	2,722	2,946	2,918	1,554	1,392	1,588	1,330
Fixed effects included in model	Grade-subject	District-grade-subject	District-subject		District-grade		District	District	District	District
Interaction terms included in model		Test-grade-subject	Test-subject		Test-grade		—	—	—	—

Note. All models are weighted by the inverse of the sampling variance of the gender gap. Standard errors (in parentheses) are clustered by state. The models include data from 2008–2009 ED*Facts* and Northwest Evaluation Association (NWEA) data sources from Grades 4 and 8. The models are restricted to district by grade cells with gap data from both ED*Facts* and NWEA. Model 1 and Model 2 are identical except that Model 1 adds the proportion of short and extended response items together while Model 2 estimates coefficients for these measures separately. Both models also include the proportion of "other" (not shown) items.
$^{+}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

than in math, we cannot reject the null hypothesis that the true association is the same in each grade and subject.

Discussion

We find that measured gender gaps are more male-favoring on state accountability tests on which a larger proportion of the overall score is based on multiple-choice items compared with tests on which a larger proportion of the overall score is based on constructed-response items. This association holds even when we control for the state gender achievement gap on a second test that has the same format and content in all states. Although the association appears smaller on math tests than reading tests, we cannot reject the hypothesis that the association is the same across subjects and grades.

These results suggest that if students are assessed using tests that weight multiple-choice questions heavily in students' total scores, the measured male-female achievement gap will favor male students more than on tests that weight constructed-response items more heavily. Although we cannot determine the reasons for the difference in measured gender gaps on tests with different item formats, our findings suggest the differences are large enough to have meaningful consequences for students. We find that on average, the gender achievement gap favors male students by one-tenth of a standard deviation more on tests with 100% multiple-choice items compared to tests with 50% constructed-response items. To give a sense of the practical meaning of a tenth of a standard deviation difference in measured gaps, suppose male and female students had the same true (normal) distribution of some set of skills. If all males' test scores were increased by one-tenth of a standard deviation, then they would make up roughly 55% of the top 10% of the observed distribution and females only 45%.

Our findings are consistent with earlier research suggesting that measured gender gaps are sensitive to the item format on standardized tests (Lindberg et al., 2010; Taylor & Lee, 2012). However, our research design has several methodological advantages over earlier work in this area. Because we use state accountability test data from 47 states, our analysis has broad generalizability to the

kinds of high-stakes accountability tests used in the United States. In addition, our use of a second test to control for between-state or -district differences in gender gaps on a common test increases the likelihood that our findings are not biased by any correlation between test item format and the magnitude of gender gaps in different populations.

One limitation of our study is that we use test data from 2008–2009. To the extent that test content and item formats have changed, particularly with the advent of tests aligned with the Common Core State Standards and other new state standards, our results may not generalize to some of the tests being used for accountability purposes today. An additional limitation is that we—like most prior studies of this issue—cannot determine whether the association between item format and gender gaps results from consistently measured gender differences in the set of math and ELA skill constructs tested by different types of items or gender differences in ancillary, construct-irrelevant skills that differentially affect performance on items of different types (e.g., handwriting skills or willingness to guess). To answer that question, researchers will need access to the content and format of each item on state accountability tests as well as item-level student response data, similar to that used by Taylor and Lee (2012). A third limitation stems from the fact that given the data available, we can only measure achievement gaps as differences in the means of the male and female score distributions. Without more detailed data, we cannot determine whether the patterns we observe are constant across the range of student achievement.

Standardized state accountability tests are used by many school districts to assign students to courses. The evidence, then, that *how* male and female students are tested (with multiple-choice or constructed-response questions) changes the perception of their relative ability in both math and ELA suggests that we must be concerned with questions of test fairness and validity: Does the assessment measure the intended skills? Does the assessment produce consistent scores for different student subgroups? And is the assessment appropriate for its intended use (Caines, Bridglall, & Chatterji, 2014; Camilli, 2006; Kane, 2013; Xi, 2010)? If the item format–related differences in apparent academic skill arise because of construct-irrelevant gender differences in average responses to multiple-choice and constructed-response items, then at least some of the standardized tests used by states necessarily fail these fairness and validity criteria.

Whether and where tests unfairly privilege male students or unfairly privilege female students, we cannot say, however. To answer this question and to determine what test developers and educators should do in response, it is essential to investigate whether construct-relevant or -irrelevant skill differences account for these patterns. If the association of gender gaps and item format is driven by differences in average levels of ancillary, construct-irrelevant skills, we need studies that can sharply identify the construct-irrelevant skills driving the patterns of difference, and test developers will need to redesign items based on this evidence.

If, however, the association of gender gaps and item format is driven by gender differences in average construct-relevant skills, this implies that the construct of interest is multidimensional and that gender gaps differ among the underlying dimensions being measured. Males' and females' relative average performance on a test and the validity of the test for its intended purpose will depend on the mix of dimensions reflected in the items on the test. Two tests measuring the same underlying constructs may rank males and females differently in performance depending on the mix of items on the test. In this case, it is particularly imperative that tests are designed to weight in appropriate proportions the mix of skills identified in states' standards.

In either case, the wide variation among states in the item format of tests indicates that where a student lives affects his or her measured performance on standardized high-stakes assessments relative to members of the other gender. This implies that test developers and educators will need to attend more carefully to the mix of item types and the multidimensional set of skills measured by tests to be sure they provide fair and appropriate measures of academic skills for both male and female students. Policymakers, too, will need to be aware of how states' use of different test formats or emphases on different skills may influence cross-state comparisons of gender gaps and funding decisions based on those results.

NOTES

This research was supported by grants from the Institute of Education Sciences (R305D110018 and R305B090016), Spencer Foundation (Award No. 201500058), and William T. Grant Foundation (Award No. 186173) to Stanford University (Sean F. Reardon, Principal Investigator). We thank Joseph Van Matre for excellent research assistance. Some of the data used in this paper were provided by the National Center for Education Statistics (NCES); additional data were provided by the Northwest Evaluation Association (NWEA). The opinions expressed here are ours and do not represent views of NCES, Institute of Education Sciences, U.S. Department of Education, or NWEA.

[1]There is a rich literature exploring the use of multiple-choice and constructed-response questions within a single test. Of particular relevance here is research on whether multiple-choice and constructed-response questions are construct equivalent and how they should be weighted into the total score on a test (Rodriguez, 2003; Wainer & Thissen, 1993). In a meta-analysis, Rodriguez (2003) finds evidence that multiple-choice and constructed-response items can be designed to be highly correlated. When they are used to test different content areas of the tested construct or draw on other cognitive skills (e.g., in essay writing), however, the correlations between the items of different types are lower. Although this work does not speak to gender differences by item type, it does suggest that there may be overall differences in performance on items of different types, particularly when these items measure different constructs or draw on ancillary skills.

[2]For the 2008–2009 school year, data are available for all states and grades except for Louisiana (which did not report test score data disaggregated by gender), California, Virginia (where eighth-grade math data cannot be used because not all students took the same state math tests in Grade 8 in 2008–09), Nebraska (where each district administered its own tests in 2008–2009), Colorado, and Florida (where only two proficiency categories were reported in 2008–2009, meaning that achievement gaps cannot be accurately computed).

[3]The reports also provide the proportion of items of each format. We prefer the proportion of score because it weights items by their contribution to the score and therefore their contribution to the gender gap. However, in supplementary analyses (not shown), we find the results are similar, albeit less precise, when we use the proportion of items as the key variable instead.

[4]The National Assessment of Educational Progress (NAEP) reading assessment is used as a proxy of English language arts (ELA) because the NAEP report that is used analyzes how each state's "reading standards for proficient performance at grades 4 and 8 in 2009 map onto the NAEP scale." NAEP uses two assessment types interchangeably because assessments in Grades 4 and 8 include a mix of reading and ELA content (National Center for Education Statistics, 2011).

[5]We use nonparametric methods for computing V from the raw (continuous) NAEP and Northwest Evaluation Association (NWEA) student test scores and maximum likelihood to estimate V from the coarsened Ed*Facts* data (see Reardon & Ho, 2015).

[6]To see this, suppose that Y_{it}, the test score of student i on test t, depends on the student's measured academic achievement (A_i), the proportion of score from non–multiple choice items on the test (p_t), some other (potentially unobserved) feature(s) of the test (U_t), and the interaction of p_t and U_t) with a student's gender (M_i, where M_i =1 for males and $M_i = 0$ for females):

$$Y_{it} = A_i + \beta_1 p_t + \beta_2 U_t + \delta\left(p_t \cdot M_i\right) + \eta\left(U_t \cdot M_i\right) + e_{it};$$
$$e_{it} \perp M, A.$$

Then the male-female gap on test t will be:

$$\begin{aligned} G_t &= E\left[Y_{it}|M_i = 1\right] - E\left[Y_{it}|M_i = 0\right] \\ &= E\left[A_i|M_i = 1\right] - E\left[A_i|M_i = 0\right] + \delta\, p_t + \eta U_t + \\ &\quad E\left[e_{it}|M_i = 1\right] - E\left[e_{it}|M_i = 0\right] \\ &= \gamma + \delta\, p_t + \eta U_t. \end{aligned}$$

[7]Drawing this conclusion depends on there being no other differences between the two tests that differentially affect males' and females' performance. If, for example, the state test was given later in the year than NAEP or emphasized different content than NAEP or if there were gender differences in how much effort males and females put into the high-stakes state tests versus the low-stakes NAEP test, we would not be able to determine whether the difference in gaps between the tests were due to the difference in item formats, a change in the gender gap over time, differences in gender gaps across different content areas, or gender differences in effort on high- and low-stakes tests.

[8]Note that Equation 2 is equivalent to a regression of the difference in the test a and n gaps within state/district s on the proportion of total score from non–multiple choice items on test a:

$$\Delta_s = G_{sa} - G_{sn} = \alpha + \delta\, p_{sa} + u_s,$$

where $u_s = u_{sa} - u_{sn}$. The estimate of δ from this model will be unbiased if the difference in the sources of error in gender gaps between the two tests is uncorrelated with p_{sa}. If we assume that u_{sn} is constant across states/districts (because the NAEP and NWEA tests are the same across places), then the model will produce an unbiased estimate of δ if $p_{sa} \perp u_{sa}$.

REFERENCES

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*(3), 219–234.

Abedi, J., Lord, C., & Plummer, J. R. (1997). *Final report of language background as a variable in NAEP mathematics performance*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Beller, M., & Gafni, N. (2000). Can item format (multiple-choice vs. open-ended) account for gender differences in mathematics achievement? *Sex Roles*, *42*(1–2), 1–21.

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.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement*, *27*(2), 165–174.

Caines, J., Bridglall, B. L., & Chatterji, M. (2014). Understanding validity and fairness issues in high-stakes individual testing situations. *Quality Assurance in Education*, *22*(1), 5–18.

Camilli, G. (2006). Test fairness. *Educational Measurement*, *4*, 221–256.

Chatterji, M. (2006). Reading achievement gaps, correlates, and moderators of early reading achievement: Evidence from the Early Childhood Longitudinal Study (ECLS) kindergarten to first grade sample. *Journal of Educational Psychology*, *98*(3), 489–507.

DeMars, C. E. (1998). Gender differences in mathematics and science on a high school proficiency exam: The role of response format. *Applied Measurement in Education*, *11*(3), 279–299.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*(1), 55–77.

Dimitrov, D. M. (1999). *Mathematics and science achievement profiles by gender, race, ability, and type of item response*. Retrieved from https://files.eric.ed.gov/fulltext/ED431788.pdf

Fryer, R. G., Jr., & Levitt, S. D. (2009). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics*, *2*(2), 210–240.

Gamer, M., & Engelhard, G., Jr. (1999). Gender differences in performance on multiple-choice and constructed-response mathematics items. *Applied Measurement in Education*, *12*(1), 29–51.

Hastedt, D., & Sibberns, H. (2005). Differences between multiple choice items and constructed response items in the IEA TIMSS surveys. *Studies in Educational Evaluation, 31*(2–3), 145–161.

Ho, A. D. (2009). A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics*, *34*(2), 201–228.

Ho, A. D., & Haertel, E. H. (2006). *Metric-free measures of test score trends and gaps with policy-relevant examples* (CSE Report 665). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Ho, A. D., & Reardon, S. F. (2012). Estimating achievement gaps from test scores reported in ordinal "proficiency" categories. *Journal of Educational and Behavioral Statistics*, *37*(4), 489–517.

Husain, M., & Millimet, D. L. (2009). The mythical "boy crisis"? *Economics of Education Review*, *28*(1), 38–48.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in math performance: A meta-analysis. *Psychological Bulletin*, *107*(2), 139–155.

Kane, M. (2013). Validity and fairness in the testing of individuals. In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability and equity* (pp. 17–53). Bingley, UK: Emerald.

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.

Lee, J., Moon, S., & Hegar, R. L. (2011). Mathematics skills in early childhood: Exploring gender and ethnic patterns. *Child Indicators Research*, *4*(3), 353–368.

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, *136*(6), 1123–1135.

Mullis, I. V. S., Martin, M. O., Fierros, E. G., Goldberg, A. L., & Stemler, S. E. (2000). *Gender differences in achievement*. Boston,

MA: International Study Center, Lynch School of Education, Boston College.
National Center for Education Statistics. (2009a). *The nation's report card: 2009 Grade 4 sample questions for mathematics, reading, and science*. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/demo_booklet/09SQ-O-G04-MRS.pdf
National Center for Education Statistics. (2009b). *The nation's report card: 2009 Grade 8 sample questions for civics, geography, U.S. history, mathematics, reading and science probe*. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/demo_booklet/09SQ-G08-MRS.pdf
National Center for Education Statistics. (2011). *A profile of state assessment standards: 2009*. Retrieved from http://nces.ed.gov/nationsreportcard/studies/statemapping/profile_standards_2009.aspx
O'Neil, H. F., Jr., & Brown, R. S. (1998). Differential effects of question formats in math assessment on metacognition and affect. *Applied Measurement in Education*, *11*(4), 331–351.
Penner, A. M., & Paret, M. (2008). Gender differences in mathematics achievement: Exploring the early grades and the extremes. *Social Science Research*, *37*(1), 239–253.
Reardon, S. F., & Ho, A. D. (2015). Practical issues in estimating achievement gaps from coarsened data. *Journal of Educational and Behavioral Statistics*, *40*(2), 158–189.
Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal*, *48*(2), 268–302.
Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, *40*(2), 163–184.
Roe, A., & Taube, K. (2003). Reading achievement and gender differences. In S. Lie, P. Linnakyla, & A. Roe (Eds.), *Northern lights on Pisa: Unity and diversity in the Nordic countries in PISA 2000* (pp. 21–36). Oslo, Norway: University of Oslo.
Routitsky, A., & Turner, R. (2003). *Item format types and their influence on cross-national comparisons of student performance*. Paper presented at the Annual Meeting of the American Educational Research Association.
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.
Sohn, K. (2012). A new insight into the gender gap in math. *Bulletin of Economic Research*, *64*(1), 135–155.
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.
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.
Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*(2), 103–118.
Willingham, W. W., & Cole, N. S. (2013). *Gender and fair assessment*. London: Routledge.
Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–170.
Zhang, L., & Manon, J. (2000). *Gender and achievement—Understanding gender differences and similarities in mathematics assessment*. Retrieved from https://eric.ed.gov/?id=ED443828

AUTHORS

SEAN F. REARDON, EdD, is the endowed Professor of Poverty and Inequality in Education at Stanford University, CERAS building, 520 Galvez Mall, #526, Stanford, CA 94305-3084; *sean.reardon@stanford.edu*. His research focuses on educational opportunity and inequality in the United States.

DEMETRA KALOGRIDES, PhD, is a research associate at the Center for Education Policy Analysis at Stanford University, 520 Galvez Mall Drive, Stanford, CA 94305; *dkalo@stanford.edu*. Her research focuses on achievement gaps, segregation, and teacher and principal labor markets.

ERIN M. FAHLE, MS, is a doctoral candidate at the Stanford University Graduate School of Education, 520 Galvez Mall, Fifth Floor, Stanford, CA 94305; *efahle@stanford.edu*. Her research focuses on using quantitative methods to understand the causes and consequences of systemic gender, racial, and economic inequalities in U.S. education.

ANNE PODOLSKY, JD, MA, is a researcher at the Learning Policy Institute, 1530 Page Mill Road, Suite 200, Palo Alto, CA 94304; *apodolsky@learningpolicyinstitute.org*. Her research focuses on improving educational opportunities and outcomes, especially for students from underserved communities.

ROSALÍA C. ZÁRATE is a doctoral candidate in developmental and psychological sciences at the Stanford Graduate School of Education, part of the Center for Education Policy Analysis Labs, and completing the Stanford master's in statistics program, 520 Galvez Mall, Stanford, CA 94305; *rzarate@stanford.edu*. Her mixed methods research focuses on higher education policy, retaining underrepresented students in higher education, and improving equity in the STEM (science, technology, engineering, and mathematics) fields.

Manuscript received September 6, 2016
Revisions received October 24, 2017,
and January 25, 2018
Accepted January 28, 2018