Test Format & Gender Achievement Gaps in the U.S.

Sean F. Reardon, Demetra Kalogrides, Erin M. Fahle, Anne Podolsky, Rosalia Zarate

Background

Standardized test construction takes measures to ensure that test items are not biased against any group of students. However, the methods used to remove biased items may fail to identify bias due to test format – the mode of student response. Several studies suggest that males perform better relative to females on multiple-choice items, and females perform better relative to males on short answer and extended-response questions, though the evidence is not conclusive (Lindberg, Hyde, Petersen, & Linn 2010; Beller & Gafni, 2000; Gamer & Engelhard, 1999; DeMars, 1998; Bolger & Kellaghan, 1990).

State accountability test formats vary significantly across the U.S. (Figure 1). Given the evidence that test format may lead to gender bias, it is unclear to what extent variable item formats affect the measurement of gender achievement gaps and "distort" comparisons across states.

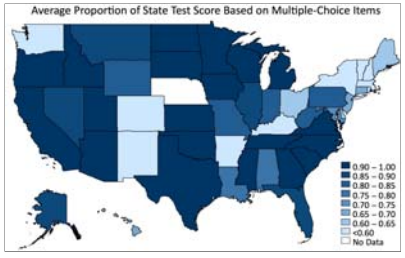


Figure 1

Current Study

Using the test scores of roughly eight million students tested in the fourth and eighth grades in English Language Arts (ELA) and math during the 2008-09 school year, we estimate state- and district-level math and ELA gender achievement gaps on each state's accountability tests, as well as on "audit" tests that do not vary in format across states and districts. We then characterize the extent to which the variation of gender achievement gaps on standardized tests across the U.S. can be explained by differing state accountability test formats.

Data

We use 4th and 8th grade math and reading achievement data in the 2008-09 school year.

EDFacts State Accountability Test Data: district-level achievement data for all students in all states on their respective state accountability tests, which vary in format across states.

National Assessment for Educational Progress (NAEP) State Data: student-level achievement data in math and reading from a representative sample of students in every state. The format of the NAEP test is the same across states.

Northwest Evaluation Association Measures of Academic Progress (MAP) Data: student-level achievement data for all students in districts that purchase the MAP assessment. The MAP test format is 100% multiple-choice.

Methods

We compare gender gaps on tests with different variable item structures that are administered to the same population of students. Specifically, we compare gaps on the state assessments with gaps on a test with a constant format across states/districts (NAEP or MAP). In this framework, a systematic association between measured gender gaps and the item structure suggests that item format may differentially affect male and female students' scores.

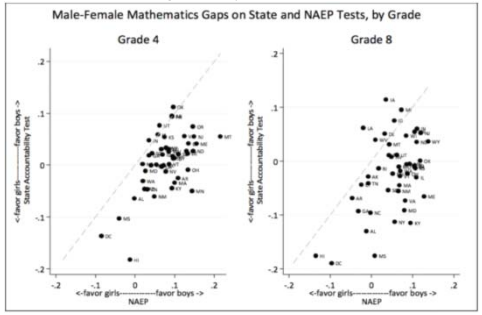
We estimate gender achievement gaps with the V-statistic (Ho & Reardon, 2012; Reardon & Ho, 2015). We then model the association as:

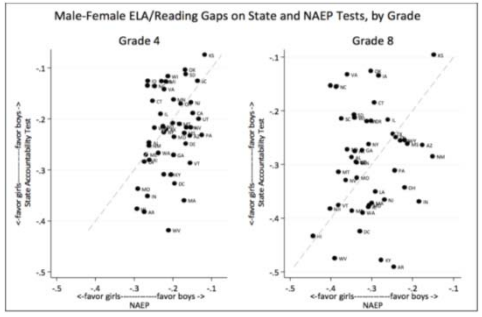
$$G_{st} = \gamma_s + \delta p_{st} + u^*_{st}$$

$$= \gamma_s + (\delta p_{sa})T_{st} + (\delta p_n)(1 - T_{st}) + u^*_{st}$$

where G_{st} is the male-female achievement gap on test t in state/district s, p_{st} is the proportion of score based on constructed response items on test t in state s; γ_s is the state gap that would be observed if a test had only multiple-choice items; and u^*_{st} is any other gender bias. The state test is denoted as a, and the national test as n. T_{st} is an indicator for the state test.

Figures 2 (top) & 3 (bottom)





Results

On the NAEP math test, males outperform females in nearly all states, whereas females outperform males in almost half of states in both grades on the state math assessments (Figure 2). In reading, the gender achievement gaps on the NAEP and state assessments again do not appear highly correlated however, they both are negative, indicating that females outperform males in both grades (Figure 3).

We find that increasing the proportion of score based on constructed-response questions on an assessment from 0 to 50 percent (roughly the largest difference between states' tests) would shift the gender achievement gap in favor of females by a little more than 0.1 standard deviations (Table 1, Model 1). There are no significant differences in the extent to which the proportion of score based on short and extended-response questions contribute to this effect (Table 1, Model 2). The consistency of the state and district results add robustness to the hypothesis that these differences are indeed driven by test format.

Test format clearly matters for the measurement of gender achievement gaps, and these differences are large enough that they may have meaningful consequences for students. However it does not determine whether females are unfairly disadvantaged by multiple-choice questions, or whether males are unfairly disadvantaged by constructed-response questions. More research is needed to understand why these differences by tests format are observed, and whether they represent real differences in skills or construct-irrelevant features of tests.

Table 1: Relationship between Proportion of Score from Multiple Choice Items on State Tests and the Size of Gender Gaps

	State-Level					*District Level*				
	Pooled Across Grades and Subjects	Math		ELA		Pooled Across Grades and Subjects	Math		ELA	
		Grade 4	Grade 8	Grade 4	Grade 8		Grade 4	Grade 8	Grade 4	Grade 8
Model 1										
Proportion Short Response+	-0.208***	-0.135**	-0.109	-0.223*	-0.376**	-0.241**	-0.126	-0.151*	-0.296**	-0.389***
Extended Response	(0.050)	(0.041)	(0.075)	(0.084)	(0.113)	(0.069)	(0.122)	(0.068)	(0.098)	(0.101)
P(Grade/Subject Interactions Are Equal)	.118					.060				
Model 2										
Proportion Short Response	-0.146*	-0.078	-0.067	-0.177+	-0.264*	-0.248**	-0.159	-0.160	-0.358	-0.326**
	(0.064)	(0.061)	(0.137)	(0.094)	(0.122)	(0.087)	(0.103)	(0.134)	(0.213)	(0.107)
Proportion Extended Response	-0.264***	-0.178***	-0.141	-0.267*	-0.5***	-0.234*	-0.088	-0.144	-0.246+	-0.437***
	(0.050)	(0.044)	(0.135)	(0.102)	(0.115)	(0.099)	(0.178)	(0.104)	(0.128)	(0.109)
P(Short Response=Extended Response)	.127	.189	.749	.391	.090	.913	.625	.936	.684	.259
N	382	96	94	96	96	5746	1530	1380	1534	1302

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.
Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. Journal of Educational Measurement, 27(2), 165-174.
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.
Gamer, M., & Engelhard Jr, G. (1999). Gender differences in performance on multiple-choice and constructed-response mathematics items. Applied Measurement in Education, 12(1), 29-51.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.
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.
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.

Acknowledgements: This research was supported by a grant from the Institute of Education Sciences (R305B090016) to the Board of Trustees of the Leland Stanford Junior University and a grant from the Spencer Foundation.