

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

CITATION: Shear, B. R. (2023). Gender bias in test item formats: Evidence from PISA 2009, 2012, and 2015 math and reading tests. *Journal of Educational Measurement*. <https://doi.org/10.1111/jedm.12372>

BACKGROUND: This paper investigates whether there are differences in male and female student performance on multiple-choice (MC) versus constructed-response (CR) items used on the large-scale math and reading tests. Such differences have been observed previously but have not been evaluated among recent samples of students.

DATA: The paper uses data from three recent administrations of the Programme for International Student Assessment (PISA) tests. The data include nationally representative samples of high school-age students who completed math and reading literacy tests across 35 countries, including the US, in 2009, 2012, and 2015.

RESEARCH DESIGN: The paper uses differential item functioning (DIF) analyses to test for systematic differences in how male and female students responded to CR versus MC items, after adjusting for overall differences in math and reading literacy.

KEY FINDINGS:

- Consistent with prior literature, on average male students tended to earn relatively higher scores on MC items, while female students tended to earn relatively higher scores on CR items.
- The differences were larger on average in reading than in math. This pattern was consistent across countries, although the magnitude of the differences varied by country.
- The magnitude of the differences would be large enough to substantively impact conclusions about gender differences depending on whether MC or CR items were used on a test.
 - Among the US samples, for example, switching from an entirely MC test to an entirely CR test could decrease male student scores (relative to female students scores) by 0.08 to 0.11 standard deviations on average.

CONCLUSION: The results of this study provide consistent evidence that among recent samples of high school-aged students participating in PISA, on average male students tend to earn relatively higher scores on MC test items whereas female students tend to earn relatively higher scores on CR test items. The item formats used in many prominent large-scale assessment programs vary considerably, making these differences relevant for research, policy, and practice. The design of the study does not allow strong conclusions about the cause of these differences; however, the consistency of the differences warrants attention and further study. Test developers have access to the most detailed data about student performance and are thus well-positioned to investigate these differences. Policymakers, researchers, and other audiences using test scores to compare student achievement across gender groups should consider the types of item formats used on tests when interpreting results.