DIVERSITY

Gender Similarities Characterize Math Performance

Standardized tests in the U.S. indicate that girls now score just as well as boys in math.

Janet S. Hyde,[1]* Sara M. Lindberg,[1] Marcia C. Linn,[2] Amy B. Ellis,[3] Caroline C. Williams[3]

Gender differences in mathematics performance and ability remain a concern as scientists seek to address the underrepresentation of women at the highest levels of mathematics, the physical sciences, and engineering (*1*, *2*). Stereotypes that girls and women lack mathematical ability persist and are widely held by parents and teachers (*3*–*5*).

Meta-analytic findings from 1990 (*6*, *7*) indicated that gender differences in math performance in the general population were trivial, $d = -0.05$, where the effect size, d, is the mean for males minus the mean for females, divided by the pooled within-gender standard deviation. However, measurable differences existed for complex problem-solving beginning in the high school years ($d = +0.29$ favoring males), which might forecast the underrepresentation of women in science, technology, engineering, and mathematics (STEM) careers.

Since this study of data from the 1970s and 1980s, several crucial cultural shifts have occurred that merit a new analysis of gender and math performance. In previous decades, girls took fewer advanced math and science courses in high school than boys did, and girls' deficit in course taking was one of the major explanations for superior male performance on standardized tests in high school (*8*). By 2000, high school girls were taking calculus at the same rate as boys, although they still lagged behind boys in the number of them taking physics (*9*). Today, women earn 48% of the undergraduate degrees in mathematics, although gender gaps in physics and engineering remain large (*10*).

Contemporary state assessments. State assessments of cognitive performance provide a contemporary source of data on these questions. Many states have conducted assessments for years, but with the advent of No Child Left Behind (NCLB) legislation, all states are mandated to conduct such assessments annually. This testing provides an exceptional opportunity to analyze current gender differences in math performance, particularly because of the extraordinary number of test takers.

Although NCLB requires states to post test results publicly, few states report data by gender and, of those that do, fewer report the necessary statistical information to compute effect sizes. Therefore, we contacted the state departments of education of all 50 states, requesting detailed statistical information on gender differences, by grade level and by ethnicity. Responses with adequate statistical information were received from 10 states: California, Connecticut, Indiana, Kentucky, Minnesota, Missouri, New Jersey, New Mexico, West Virginia, and Wyoming. In all cases, the data represent the testing of all students attending school in that grade. These states are geographically diverse and appear to be representative of all 50 states insofar as their average scores on the National Assessment of Educational Progress (NAEP, a federal assessment that carefully samples students nationwide) match the average for all 50 states quite closely. For 8th-graders, the average NAEP mathematics score was 280.22 for our 10 states and 280.17 for all 50 states (*11*).

Grade	$d \pm$ SE	Variance ratio	*N*
Grade 2	0.06 ± 0.003	1.11	460,980
Grade 3	0.04 ± 0.002	1.11	754,894
Grade 4	−0.01 ± 0.002	1.11	763,155
Grade 5	−0.01 ± 0.002	1.14	929,155
Grade 6	−0.01 ± 0.002	1.14	886,354
Grade 7	−0.02 ± 0.002	1.16	898,125
Grade 8	−0.02 ± 0.002	1.21	837,979
Grade 9	−0.01 ± 0.003	1.14	608,229
Grade 10	0.04 ± 0.003	1.18	619,591
Grade 11	0.06 ± 0.003	1.17	446,381

Effect sizes across grades for U.S. mathematics tests; results are similar across grades 2 through 11.

Percentage of children scoring above indicated percentile and ratios						
Ethnic group	Above 95th percentile			Above 99th percentile		
	F	M	M/F	F	M	M/F
Asian/Pacific Islander (*n* =219)	5.71	6.27	1.09	1.37	1.25	0.91
White (*n* = 3473)	5.38	7.80	1.45	0.90	1.85	2.06

The upper tail. Percentage of Minnesota children scoring above the 95th and 99th percentiles in 11th grade mathematics testing, by gender and ethnicity. Too few students scored above the 95th percentile to compute reliable statistics for these groups: American Indians, Hispanics, and Black not Hispanic.

Gender and average performance. Effect sizes for gender differences, representing the testing of over 7 million students in state assessments, are uniformly <0.10, representing trivial differences (see table, top left, and table S1). Of these effect sizes, 21 were positive, indicating better performance by males; 36 were negative, indicating better performance by females; and 9 were exactly 0. From this distribution of effect sizes, we calculate that the weighted mean is 0.0065, consistent with no gender difference (see chart on p. 495 and fig. S1). In contrast to earlier findings, these very current data provide no evidence of a gender difference favoring males emerging in the high school years; effect sizes for gender differences are uniformly <0.10 for grades 10 and 11 (see table, top left, and table S1). Effect sizes for the magnitude of gender differences are similarly small across all ethnic groups (table S2). The magnitude of the gender difference does not exceed $d = 0.04$ for any ethnic group in any state.

Gender and variance. Another explanation for the underrepresentation of women at the highest levels in STEM careers has focused not on averages, but on variance, the extent to which scores of one gender or the other vary from the mean score. The hypothesis that the variability of intellectual abilities is greater among males than among females and produces a preponderance of males at the highest levels of performance was originally proposed over 100 years ago (*12*).

The variance ratio (VR), the ratio of the male variance to the female variance, assesses

[1]Department of Psychology, University of Wisconsin, 1202 West Johnson Street, Madison, WI 53706, USA. [2]Education in Mathematics, Science, and Technology, University of California, Berkeley, Berkeley, CA 94720, USA. [3]Department of Curriculum and Instruction, University of Wisconsin, Madison, WI 53706, USA.

*Author for correspondence. E-mail: jshyde@wisc.edu

Published by AAAS

these differences. Greater male variance is indicated by VR > 1.0. All VRs, by state and grade, are >1.0 [range 1.11 to 1.21 (see top table on p. 494)]. Thus, our analyses show greater male variability, although the discrepancy in variances is not large. Analyses by ethnicity show a similar pattern (table S2).

Does this greater variability translate into gender differences at the upper tail of the distribution (*13*)? Data from the state assessments provide information on the percentage of boys and girls scoring above a selective cut point. Results vary by ethnic group. The bottom table on p. 494 shows data for grade 11 for the state of Minnesota. For whites, the ratios of boys:girls scoring above the 95th percentile and 99th percentile are 1.45 and 2.06, respectively, and are similar to predictions from theoretical models. For Asian Americans, ratios are 1.09 and 0.91, respectively. Even at the 99th percentile, the gender ratio favoring males is small for whites and is reversed for Asian Americans. If a particular specialty required mathematical skills at the 99th percentile, and the gender ratio is 2.0, we would expect 67% men in the occupation and 33% women. Yet today, for example, Ph.D. programs in engineering average only about 15% women (*14*).

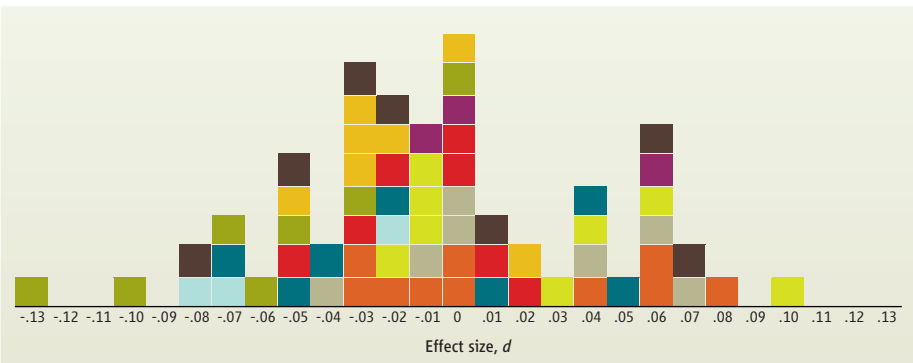
Gender and item complexity. An additional issue in assessing gender differences in math performance and the underrepresentation of women in STEM careers is the question of the cognitive complexity or depth of knowledge being tested. Earlier studies (*6*) indicated that, although girls equaled or surpassed boys in basic computation and understanding of mathematical concepts, boys exceeded girls in complex problem-solving beginning in the high school years, $d = +0.29$. Complex problem-solving is crucial for advanced work in STEM careers. At the time of the 1990 meta-analysis, girls were less likely to take advanced math and science courses, and this gender difference in course choice was a likely explanation for the gender gap in complex problem-solving (*8*).

Today, with the gender gap erased in taking advanced math courses, does the gender gap remain in complex problem-solving? To answer this question, we coded test items from all states where tests were available, using a four-level depth of knowledge framework (*15*). Level 1 (recall) includes recall of facts and performing simple algorithms. Level 2 (skill/concept) items require students to make decisions about how to approach a problem and typically ask students to estimate or compare information. Level 3 (strategic thinking) includes complex cognitive demands that require students to reason, plan, and use evidence. Level 4 (extended thinking) items require complex reasoning over an extended period of time and require students to connect ideas within or across content areas as they develop one among alternate approaches. We computed the percentage of items at levels 3 or 4 for each state for each grade, as an index of the extent to which the test tapped complex problem-solving. The results were disappointing. For most states and most grade levels, none of the items were at levels 3 or 4. Therefore, it was impossible to determine whether there was a gender difference in performance at levels 3 and 4.

The dearth of level-3 or level-4 items in state assessments has an additional serious consequence. With the increased emphasis on testing associated with NCLB, more teachers are gearing their instruction to the test (*16*). If the tests do not assess the sorts of reasoning that are crucial to careers in STEM disciplines, then these skills may be neglected in instruction, putting American students at a disadvantage relative to those in other countries where tests and curricula emphasize more challenging content (*17*).

To address this limitation in the state assessments, we returned to the NAEP data (*18*). NAEP categorizes items as easy, medium, or hard. We coded hard sample items for depth of knowledge. No items were at level 4 but many were at level 3. We computed the magnitude of gender differences on the hard items that were at level 3 depth of knowledge. At grade 12, effect sizes for these items ranged between 0 and 0.15 (average $d = 0.07$). At grade 8, effect sizes for these items ranged between 0 and 0.08 (average $d = 0.05$). Thus, even for difficult items requiring substantial depth of knowledge, gender differences were still quite small.

Conclusion. Our analysis shows that, for grades 2 to 11, the general population no longer shows a gender difference in math skills, consistent with the gender similarities hypothesis (*19*). There is evidence of slightly greater male variability in scores, although the causes remain unexplained. Gender differences in math performance, even among high scorers, are insufficient to explain lopsided gender patterns in participation in some STEM fields. An unexpected finding was that state assessments designed to meet NCLB requirements fail to test complex problem-solving of the kind needed for success in STEM careers, a lacuna that should be fixed.



Effect sizes across grades and U.S. states. The weighted mean is 0.0065, consistent with no gender difference. Each square represents the effect size for one grade within one state. New Mexico (pea green), Kentucky (pink), Wyoming (dark brown), Minnesota (teal), Missouri (red), West Virginia (gold), Connecticut (tan), California (orange), Indiana (yellow), New Jersey (purple).

References and Notes

1. National Academy of Sciences, *Beyond Bias and Barriers: Fulfilling the Potential of Women in Academic Science and Engineering* (National Academies Press, Washington, DC, 2006).
2. D. F. Halpern *et al.*, *Psychol. Sci. Public Interest* **8**, 1 (2007).
3. P. M. Frome, J. S. Eccles, *J. Pers. Soc. Psychol.* **74**, 435 (1998).
4. A. Furnham *et al.*, *J. Genet. Psychol.* **163**, 24 (2002).
5. Q. Li, *Educ. Res.* **41**, 63 (1999).
6. J. S. Hyde, E. Fennema, S. Lamon, *Psychol. Bull.* **107**, 139 (1990).
7. General guidelines are that $d = 0.20$ is a small effect, $d = 0.50$ is moderate, and $d = 0.80$ is large (*20*).
8. J. L. Meece *et al.*, *Psychol. Bull.* **91**, 324 (1982).
9. NSF, *Science and Engineering Indicators 2006,* www.nsf.gov/statistics/seind06/ (2006).
10. NSF, www.nsf.gov/statistics/wmpd/underdeg.htm (2004).
11. National Assessment of Educational Progress, http://nces.ed.gov/nationsreportcard/nde/statecomp/ (2008).
12. S. A. Shields, *Am. Psychol.* **30**, 739 (1975).
13. L. V. Hedges, L. Friedman, *Rev. Educ. Res.* **63**, 94 (1993).
14. J. Handelsman *et al.*, *Science* **309**, 1190 (2005).
15. N. L. Webb, *Appl. Meas. Educ.* **20**, 7 (2007).
16. W. Au, *Educ. Res.* **36** 258 (2007).
17. K. Roth, H. Garnier, *Educ. Leadership* **64**, 16 (2006).
18. National Assessment of Educational Progress, http://nces.ed.gov/nationsreportcard/itmrls/startsearch.asp (2008).
19. J. S. Hyde, *Am. Psychol.* **60**, 581 (2005).
20. J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Lawrence Erlbaum, Hillsdale, NJ, 1988).

21. We thank personnel in each of the 10 responding states for providing data. Special thanks to Margaret Biggerstaff of the Minnesota Department of Education for additional analyses. This research was funded through grant REC 0635444 from NSF. Any findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

10.1126/science.1160364

Supporting Online Material
www.sciencemag.org/cgi/content/full/321/5888/494/DC1