Supporting Online Material for

Gender Similarities Characterize Math Performance

Janet S. Hyde,* Sara M. Lindberg, Marcia C. Linn, Amy B. Ellis, Caroline C. Williams

*To whom correspondence should be addressed. E-mail: jshyde@wisc.edu

Published 25 July 2008, *Science* **320**, 494 (2008)
DOI: 10.1126/science.1160364

This PDF file includes

Supporting Online Material

SOM Text

This supporting online text extends the discussion of gender similarities in state assessments of mathematics performance and provides additional detail in tables and graphs. In particular, we address the issue of whether there is a preponderance of males in the upper tail of the distribution of mathematics performance and the issue of gender differences in performance on the quantitative portion of the SAT. For an elaboration of other hypotheses regarding the underrepresentation of women in science, technology, engineering, and mathematics (STEM) fields, see Halpern *et al*. (*S1*). We also provide detailed information on the definitions for coding the depth of knowledge required for solving mathematics items.

The Issue of the Tails of the Distribution

In our Education Forum (*S2*), we presented evidence for gender similarities in mathematics performance, based on state assessments of more than 7 million students. In that forum, we showed how close the average scores for males and females are. In explaining the underrepresentation of women at the highest levels of mathematics and science achievement, however, some researchers focus not on average scores in the general population, but instead on gender ratios in the upper tail of the ability distribution, reasoning that high levels of mathematical ability are necessary for success in STEM careers. This approach shifts the argument from a discussion of average scores to a discussion of variance, the extent to which scores of one gender or the other vary from the mean score.

Theoretical models quantify the consequences of greater male variance based on the assumption that scores are normally distributed (*S3*). For example, if the effect size (*d)* = +0.05 and the variance ratio (VR) of males to females = 1.12, values representative of those found in the state assessments, then the male:female ratio at the 95th percentile of the distribution would be 1.34. That is, there will be 1.34 males for every female. At a very high cut-off point, the 99.9th percentile, indicating exceptional performance, the male:female ratio would be 2.15. Even if a particular specialty required mathematical skills at the 99.9th percentile, we would expect 68% men in the occupation and 32% women. Yet today, for example, Ph.D. programs in Engineering average only about 15% women (*S4*).

Table S1 addresses this point by examining the variance ratio, or ratio of the male variance to the female variance, for each grade in each state. Values greater than 1 indicate greater male variance.

All variance ratios in Table S1, with one exception, are greater than 1. However, none are very large and, with one exception, all lie between 1.03 and 1.35, meaning that the male variance is not markedly greater than the female variance. There is no evidence that males display larger variance in scores of a magnitude that would account for the underrepresentation of women at the highest levels of performance in STEM fields.

In the Education Forum (*S2*), Table 2 also addresses this issue of greater male variability and the gender ratio at the high end of the distribution of scores, by examining gender ratios (number of males divided by number of females) among those scoring at very high levels of these tests of mathematics performance, namely the 95%ile and 99%ile. For whites, there are 1.45 times as many boys as girls above the 95%ile in grade 11, and twice as many boys as girls above the 99%ile. For Asian Americans, however, at the 99%ile, the gender ratio is 0.91, meaning that more girls than boys scored above the 99%ile. Thus the pattern of more males than females receiving exceptional scores is not universal.

The gender ratio found in the upper tail of the distribution, of course, is sensitive to many complex factors, including the depth of knowledge required by the items and the content of the items (e.g., algebra vs. geometry). Just as gender ratios in the upper tail of the distribution vary by ethnicity in these data, so also may they vary according to other factors.

The Issue of the SAT-Math

Gender differences in performance on the SAT Mathematics test are widely publicized and contribute to the public's view that males excel in mathematics, compared with females. In 2007, males scored an average of 533 ± 114 (mean ± SD = 114) on the Mathematics portion of the SAT, compared with an average of 499 ± 111 for girls (*S5*). For many reasons, these data tell us nothing about gender differences in mathematics performance. Chief among these reasons is sampling. The SAT is taken almost exclusively by college-bound students, and even then, some college-bound students do not take it because their intended college requires some other test such as the ACT. Therefore, there is no well-defined sampling frame that would permit broader generalization. Perhaps more important is the fact that, coupled with the current trend for more females than males to attend college, the SAT is taken by more females than males. In 2007 the SAT was taken by 798,030 females but only 690,500 males (*S5*), a gap of more than 100,000 people. Assuming that SAT takers represent the top portion of the performance distribution, this surplus of females taking the SAT means that the female group dips farther down into the performance distribution than does the male

group. It is therefore not surprising that females, on average, score somewhat lower than males. The gender gap is likely in large part a sampling artifact.

This conclusion is verified by results from a study of the ACT (*S6*). It, too, is taken by a selective group of college-bound students. Traditionally, males have had a slight advantage of 0.2–0.3 points on the composite score. In 2002, two states, Colorado and Illinois, mandated the administration of the ACT to all high school students in those states. The results are shown in Figure S2. The gender gap in scores disappeared when the test was administered to all students and, in fact, a slight gap favoring females emerged. These findings support the conclusion that the male advantage on the SAT mathematics test is largely an artifact of sampling.

Coding Depth of Knowledge of Mathematics Test Items

The following provides a detailed description of the definitions of the 4 levels of depth of knowledge that were used in coding mathematics test items (*S7*, *S8*).

***Level 1* (*Recall*)** includes the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula. That is, in mathematics a one-step, well-defined, and straight algorithmic procedure should be included at this lowest level. Other key words that signify a Level 1 include "identify," "recall," "recognize," "use," and "measure." Verbs such as "describe" and "explain" could be classified at different levels depending on what is to be described and explained.

***Level 2* (*Skill/Concept*)** includes the engagement of some mental processing beyond a habitual response. A Level 2 assessment item requires students to make some decisions as to how to approach the problem or activity, whereas Level 1 requires students to demonstrate a rote response, perform a well-known algorithm, follow a set procedure (like a recipe), or perform a clearly defined series of steps. Keywords that generally distinguish a Level 2 item include "classify," "organize," "estimate," "make observations," "collect and display data," and "compare data." These actions imply more than one step. For example, to compare data requires first identifying characteristics of the objects or phenomenon and then grouping or ordering the objects. Some action verbs, such as "explain," "describe," or "interpret" could be classified at different levels depending on the object of the action. For example, if an item required students to explain how light affects mass by indicating there is a relationship between light and heat, this is considered a Level 2. Interpreting information from a simple graph, requiring reading information from the graph, also is a Level 2. Interpreting information from a complex graph that requires some decisions on what features of the graph need to be considered and how information from the graph can be aggregated is a Level 3. Caution is warranted in interpreting Level 2 as only skills because some reviewers will interpret skills very narrowly, as primarily numerical skills, and such interpretation excludes from this level other skills such as visualization skills and probability skills, which may be more complex simply because they are less common. Other Level 2 activities include explaining the purpose and use of experimental procedures; carrying out experimental procedures; making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts.

***Level 3* (*Strategic Thinking*)** requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is a Level 3. Activities that require students to make conjectures are also at this level. The cognitive demands at Level 3 are complex and abstract. The complexity does not result from the fact that there are multiple answers, a possibility for both Levels 1 and 2, but because the task requires more demanding reasoning. An activity, however, that has more than one possible answer and requires students to justify the response they give would most likely be a Level 3. Other Level 3 activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and using concepts to solve problems.

***Level 4* (*Extended Thinking*)** requires complex reasoning, planning, developing, and thinking most likely over an extended period of time. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified as a Level 2. However, if the student is to conduct a river study that requires taking into consideration a number of variables, this would be a Level 4. At Level 4, the cognitive demands of the task should be high and the work should be very complex. Students should be required to make several connections—relate ideas *within* the content area or *among* content areas—and have to select one approach among many alternatives on how the situation should be solved, in order to be at this highest level. Level 4 activities include designing and conducting experiments; making connections between a finding and related concepts and phenomena; combining and synthesizing ideas into new concepts; and critiquing experimental designs.

References and Notes

S1. D. F. Halpern, C. P. Benbow, D. C. Geary, R. C. Gur, J. S. Hyde, M. Gernsbacher, *Psych. Sci. Public Interest* **8**, 1-51 (2007).

S2. J. S. Hyde, S. M. Lindberg, M. C. Linn, A. Ellis, C. Williams, *Science* **321**, xxx (2008).

S3. L. V. Hedges, L. Friedman *Rev. Educ. Res.* **63** 94-105 (1993).

S4. J. Handelsman, N. Cantor, M. Carnes et al. *Science* **309** 1190-1191 (2005).

S5. College Board, *Total Group Profile Report* (College Board, New York, 2007); www.collegeboard.com/prod_downloads/about/news_info/cbsenior/yr2007/national-report.pdf

S6. ACT, *Gender Fairness Using the ACT* (ACT, Iowa City, IA, 2005); www.act.org/research/policymakers/pdf/gender.pdf.

S7. N. L. Webb *Appl. Meas. Educ.* **20**(10), 7-25 (2007).

S8. N. L. Webb, M. Alt, R. Ely, M. Cormier, B. Vesperman. *The Web Alignment Tool: Development, Refinement, and Dissemination* (Council of Chief State School Officers, Washington, DC, 2006); www.ccsso.org/publications/index.cfm.

S9. This research was funded through grant REC 0635444 from NSF. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank the advisory board to the project: C. Dwyer, P. Holland, N. Newcombe, and A. Petersen.

Table S1. Effect sizes for gender differences, variance ratios, and total number of students tested, by state and grade.

State	Effect size (*d*)	Variance ratio	*N*	State	Effect size (*d*)	Variance ratio	*N*
California				Missouri			
Grade 2	0.06	1.11	460,980	Grade 3	0.00	1.12	64,434
Grade 3	0.06	1.11	463,969	Grade 4	–0.00	1.09	64,979
Grade 4	–0.03	1.13	470,963	Grade 5	0.02	1.15	65,835
Grade 5	–0.01	1.16	479,802	Grade 6	–0.03	1.18	66,745
Grade 6	0.00	1.15	481,557	Grade 7	–0.05	1.18	70,328
Grade 7	–0.03	1.17	481,275	Grade 8	–0.02	1.19	72,176
Grade 8	–0.02	1.16	472,790	Grade 10	0.01	1.22	67,723
Grade 9	0.00	1.13	498,613	New Jersey			
Grade 10	0.04	1.18	409,453	Grade 5	–0.01	1.17	103,796
Grade 11	0.08	1.10	337,126	Grade 6	–0.00	1.09	103,915
Connecticut				Grade 7	0.06	1.16	107,115
Grade 3	0.04	1.13	41,558	New Mexico			
Grade 4	0.06	1.14	42,308	Grade 3	–0.00	1.03	24,685
Grade 5	–0.00	1.07	42,108	Grade 4	–0.05	1.04	24,372
Grade 6	–0.04	1.11	43,025	Grade 5	–0.03	1.09	24,758
Grade 7	–0.01	1.11	43,828	Grade 6	–0.10	1.08	25,153
Grade 8	0.00	1.15	43,944	Grade 7	–0.07	1.11	25,548
Grade 10	0.07	1.14	42,255	Grade 8	–0.06	1.08	25,987
Indiana				Grade 9	–0.13	1.17	27,691
Grade 3	0.04	1.12	77,097	West Virginia			
Grade 4	0.10	1.08	77,362	Grade 3	–0.03	1.03	19,549
Grade 5	–0.01	1.11	78,818	Grade 4	–0.03	1.15	19,657
Grade 6	–0.01	1.15	79,023	Grade 5	0.02	1.24	20,466
Grade 7	–0.01	1.17	79,905	Grade 6	–0.03	1.24	20,779
Grade 8	0.03	1.19	80,854	Grade 7	–0.02	1.24	21,112
Grade 9	–0.02	1.16	81,925	Grade 8	0.00	1.14	21,500
Grade 10	0.06	1.14	80,674	Grade 10	–0.05	1.35	19,486
Kentucky				Wyoming			
Grade 5	–0.02	1.06	48,956	Grade 3	0.06	1.03	6,199
Grade 8	–0.07	2.39	50,474	Grade 4	0.01	1.12	6,279
Grade 11	–0.08	1.76	40,669	Grade 5	0.07	1.07	6,063
Minnesota				Grade 6	–0.03	1.10	6,416
Grade 3	0.04	1.10	57,403	Grade 7	–0.08	1.10	6,669
Grade 4	–0.04	0.94	57,235	Grade 8	–0.02	1.11	6,826
Grade 5	–0.02	1.05	58,553	Grade 11	–0.05	1.30	6,060
Grade 6	–0.05	1.09	59,741				
Grade 7	0.01	1.16	62,345				
Grade 8	–0.07	1.07	63,428				
Grade 11	0.05	1.23	62,526				

Table S2. Effect sizes and variance ratios, by state and ethnicity.

State	Effect size (*d*)	Variance ratio	*N*
Connecticut			
American Indian/Alaskan Native	–0.06	1.25	1,032
Asian/Pacific Islander	0.04	1.06	10,325
Black	–0.10	1.16	40,754
Hispanic	0.02	1.16	43,984
Caucasian	0.04	1.11	202,931
Minnesota			
American Indian/Alaskan Native	–0.05	1.09	8,367
Asian/Pacific Islander	0.00	1.05	24,307
Black	–0.08	1.09	34,530
Hispanic	0.00	1.09	21,475
Caucasian	–0.00	1.09	332,552
New Mexico			
American Indian/Alaskan Native	–0.10	1.06	24,262
Asian/Pacific Islander	–0.01	1.16	2,095
Black	–0.12	1.10	4,355
Hispanic	–0.06	1.09	94,230
Caucasian	–0.06	1.10	53,252
Overall			
American Indian/Alaskan Native	–0.08	1.07	33,661
Asian/Pacific Islander	0.01	1.06	36,727
Black	–0.09	1.13	79,639
Hispanic	–0.03	1.11	159,689
Caucasian	0.01	1.10	588,735
Mixed/not reported (all other states)	0.01	1.15	6,306,392

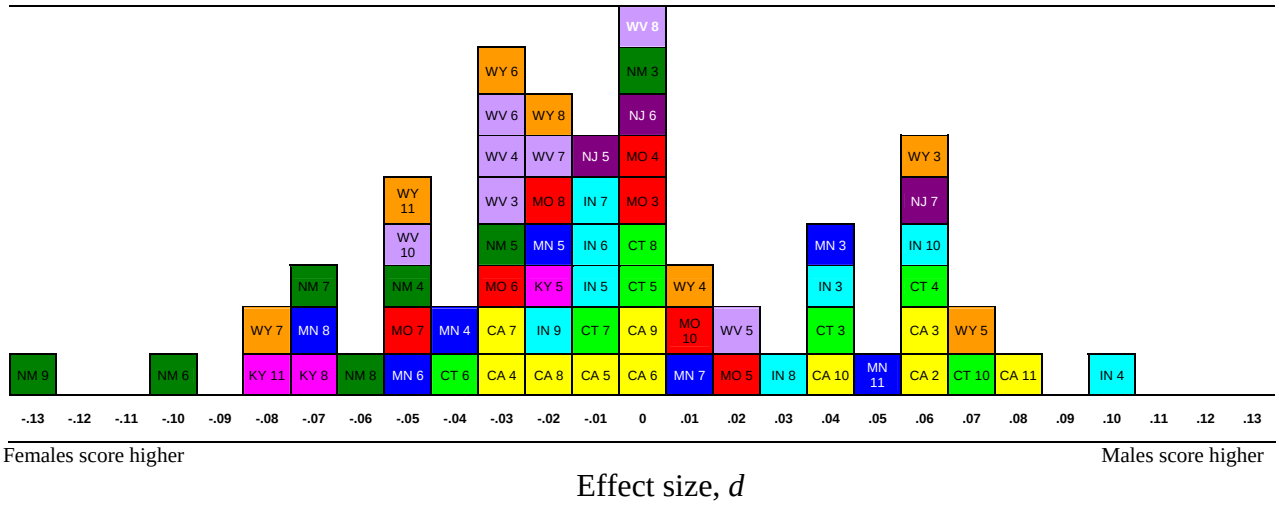

Fig. S1. Effect sizes, *d*, for each grade and each state. The weighted mean is 0.0065, consistent with no gender difference.

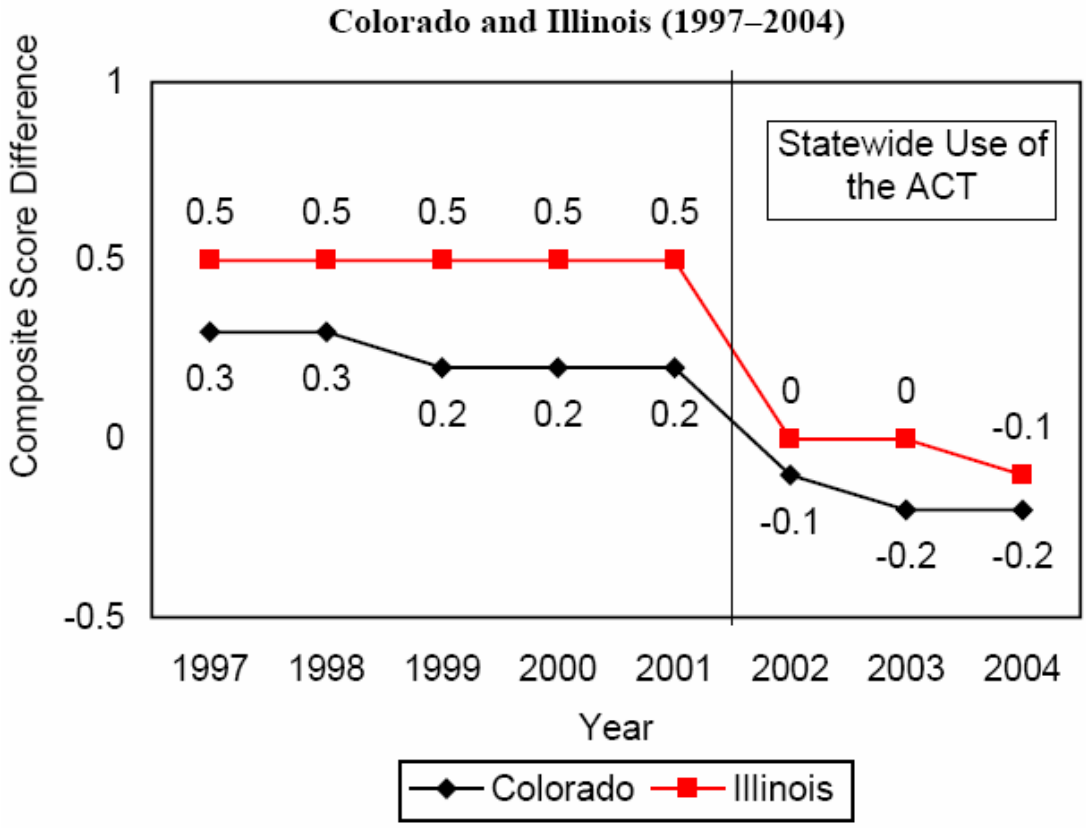

Fig. S2. Gender difference (male-female) in ACT composite scores, Colorado and Illinois, from 1997 to 2004.