Journal of Economic Perspectives—Volume 24, Number 2—Spring 2010—Pages 109–128

The Gender Gap in Secondary School Mathematics at High Achievement Levels: Evidence from the American Mathematics Competitions

Glenn Ellison and Ashley Swanson

Boys used to take substantially more math courses in high school than did girls. As the course-taking gap has narrowed, so too has the gap in averages on standardized tests. The precise size of the gap varies from test to test and country to country. While new estimates continue to attract considerable attention, most findings are qualitatively similar: a gap on math tests remains, but it is sufficiently small so as to be of little practical importance. For example, Hyde, Lindberg, Linn, Ellis, and Williams (2008) find a small to nonexistent gender gap; Freeman (2005), Perie, Moran, and Lutkus (2005), OECD (2006), Guiso, Monte, Sapienza, and Zingales (2008), and Penner (2008) find small to moderate gaps in the United States and other countries; and Fryer and Levitt (forthcoming) are on the high side, reporting that a 0.2 standard deviation gap emerges in the United States by fifth grade.

The gender gap on math tests among high-achieving students is consistently much larger. For example, there is a 2.1 to 1 male–female ratio among students scoring 800 on the math SAT, and a ratio of at least 1.6 to 1 among students scoring in the 99th percentile on the Program for International Student Assessment (PISA) test in 36 of the 40 countries studied by Guiso, Monte, Sapienza, and Zingales (2008). The existence of a gap of this magnitude by the end of high school is troubling both for reasons of gender fairness and because failures to develop the talent of any group have aggregate consequences. With respect to the education

■ *Glenn Ellison is Gregory K. Palm Professor of Economics and Ashley Swanson is a Ph.D. student in Economics, both at the Massachusetts Institute of Technology, Cambridge, Massachusetts. Ellison is also Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts and MathCounts Coach, Bigelow Middle School, Newton, Massachusetts. Their e-mail addresses are ⟨gellison@mit.edu⟩ and ⟨aswan@mit.edu⟩.*

doi=10.1257/jep.24.2.109

literature, this gap suggests that policies that serve the average girl well may not serve high-achieving girls well. It is also potentially relevant to other literatures. For example, it may relate to the underrepresentation of girls in math and science; Xie and Shauman (2003) present a nice overview of research into dozens of factors that may be important here. The lack of women in technical fields, in turn, appears to be a significant contributor to the gender gap in wages (for example, Brown and Corcoran, 1997; Blau and Kahn, 2000).

This paper presents new evidence on the gender gap in secondary school math at high achievement levels using data from the American Mathematics Competitions (AMC), a series of contests sponsored by the Mathematical Association of America. The contests are given in about 3,000 U.S. high schools and about 225,000 students participate. The primary aspect of these contests that makes them attractive as a source of research data is that they are explicitly designed to distinguish among students at very high achievement levels.

Our analysis enriches existing descriptive work on the gender gap in several ways. Most fundamentally, the AMC data provide a clearer picture of the magnitude of the gender gap at very high performance levels. Here, our most striking finding is that the gender gap appears to widen substantially at percentiles beyond the 99th: at the very high end of our data, the male–female ratio exceeds 10 to 1. (At less extreme percentiles we find that women are more underrepresented among high scorers on the AMC contests than they are among students with comparable performance on the SATs, suggesting that these contests may be less appealing to high-achieving girls.) Our second set of analyses examines whether and how the gender gap varies across schools within the United States.[1] We find some variation across schools, but the magnitude of the variation is only moderate—we estimate that there is a substantial gender gap in almost every U.S. high school, but there is enough variation from school to school to suggest that the number of girls reaching high performance levels would increase substantially if all school environments could somehow be made to resemble those where girls are currently doing relatively well. Finally, we examine extreme high-achieving students chosen to represent their countries in international competitions. Here, our most striking finding is that the highest-scoring boys and the highest-scoring girls in the United States appear to be drawn from very different pools. Whereas the boys come from a variety of backgrounds, the top-scoring girls are almost exclusively drawn from a remarkably small set of super-elite schools: as many girls come from the 20 schools that generally do best on these contests as from all other high schools in the United States combined. This suggests that almost all American girls with extreme mathematical ability are not developing their mathematical talents to the degree necessary to reach the extreme top percentiles of these contests.

[1] Previous papers including Guiso, Monte, Sapienza, and Zingales (2008), Machin and Pekkarinen (2008), Penner (2008), and Andreescu, Gallian, Kane, and Mertz (2008) have examined variation in the gender gap across countries, which may derive from cultural and other factors.

The American Gender Gap in Math Scores at Higher Percentiles

In this section, we present some data on the gender gap at high levels of the American Mathematics Competition test scores. We begin with a description of these tests and some comparisons to other tests. Our most basic finding is that the gender gap on the AMC is large and widens dramatically at very high percentiles.

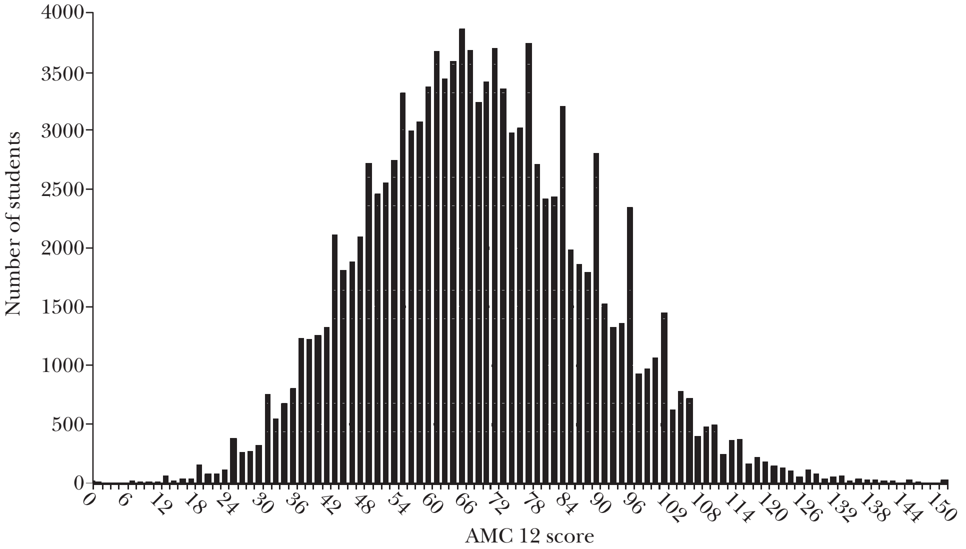
Measuring High Math Achievement: The American Mathematics Competitions and Other Tests

The Mathematical Association of America has sponsored the American Mathematics Competitions (AMC) since 1950. They are given in over 3,000 high schools and at a number of other locations. While 3,000 is a small fraction of all U.S. high schools, the AMC is much more likely to be offered in high-achieving high schools, which probably makes it available to a majority of the top students in the United States. Some schools have hundreds of students participate, but it is far more common for the AMC exams to be taken by a few dozen self-selected students. Our primary focus will be on the AMC 12, which is taken by about 125,000 students in a typical year. Test takers are roughly evenly distributed between grades eleven and twelve, with a smattering in grades ten and below. The test is offered on just two dates each year: the first, in the second week of February, is referred to as the 12A and the second, which occurs fifteen days later, is referred to as the 12B. High scorers (typically students who score above 100) are invited to participate in subsequent AMC contests. The AMC series of contests are the most prestigious high school contests in the United States and lead eventually to the selection of students to represent the United States in the International Math Olympiad. Some elite colleges, including MIT, Cal Tech, Yale, and Brown, invite students to report AMC scores on their application forms.

The AMC contests are designed to distinguish among students at high performance levels. Relative to the SAT, this is accomplished by asking fewer questions per unit time and making the questions—which are of progressive difficulty—much harder. Figure 1 is a histogram of AMC 12 scores for 2007. Scores range from zero to 150, with an average score of around 65. In several places our analyses of "high-achieving" students will focus on students who score at least 100 on the AMC 12. Our impression from casual experience and a small study we discuss in Ellison and Swanson (2009) is that scoring 100 on the AMC 12 can be thought of as roughly comparable to scoring 780–800 (the 99th percentile) on the math SAT. About 6 percent of U.S. AMC test takers scored at least 100 on the 2007 AMC 12. We will also sometimes examine students reaching higher score levels. About 0.8 percent of U.S. test takers scored at least 120. About 0.2 percent scored at least 130.

To illustrate the material being tested and the level of mastery needed to score 100 on the AMC 12, Figure 2 presents some sample questions from the 2007 AMC 12A. The full test includes 25 problems to be solved in 75 minutes. A student can achieve a score of 100 by solving 14 problems and leaving eleven blank. The questions increase in difficulty, so the sample questions, which were numbers

Figure 1
AMC 12 Score Histogram



Note: This figure is a histogram of AMC 12 scores for 2007. Scores range from zero to 150, with an average score of around 65.

13–17, were often critical in determining whether a student reached 100. Note that some of the problems require specific knowledge of precalculus high school math topics like equations for lines and trigonometric identities, whereas others mainly test problem-solving skills—for example, whether the student can formulate a strategy for answering an unusual problem and carry out the calculations in an organized manner. We have no academic credentials for making such assessments, but would say that to our untrained eyes the questions seem like good ones for assessing the math skills we would like to see in economics students. The last few questions on the AMC 12 are very difficult and are designed to distinguish among students at higher percentiles. For example, the 2007 AMC 12A included two questions answered correctly by only 20–25 percent of students who scored at least 100, and three that were answered correctly by fewer than 11 percent of such students.

We will also present some data on other AMC contests. The AMC 10 is a contest similar to the AMC 12, but is open to students in grades ten and below. It is given in most of the same schools (at the same times) and is taken by approximately 100,000 students per year. The test is also designed so that the mean score is about 65. Students who took both the AMC 10 and the AMC 12 in 2007 scored about 13 points higher on average on the AMC 10. The American Invitational Math Exam (AIME) is a more demanding contest. Students get three hours to work on 15 problems with numerical answers. It is open only to students who have achieved a qualifying score of approximately 120 on the AMC 10 or 100 on the AMC 12. The problems are

Figure 2
Questions 13 through 17 from the 2007 AMC 12A

13. A piece of cheese is located at $(12, 10)$ in a coordinate plane. A mouse is at $(4, -2)$ and is running up the line $y = -5x + 18$. At the point (a, b) the mouse starts getting farther from the cheese rather than closer to it. What is $a + b$?

    (A) 6 **(B)** 10 **(C)** 14 **(D)** 18 **(E)** 22

14. Let a, b, c, d, and e be distinct integers such that

    $$(6 - a)(6 - b)(6 - c)(6 - d)(6 - e) = 45.$$

    What is $a + b + c + d + e$?

    (A) 5 **(B)** 17 **(C)** 25 **(D)** 27 **(E)** 30

15. The set $\{3, 6, 9, 10\}$ is augmented by a fifth element n, not equal to any of the other four. The median of the resulting set is equal to its mean. What is the sum of all possible values of n?

    (A) 7 **(B)** 9 **(C)** 19 **(D)** 24 **(E)** 26

16. How many three-digit numbers are composed of three distinct digits such that one digit is the average of the other two?

    (A) 96 **(B)** 104 **(C)** 112 **(D)** 120 **(E)** 256

17. Suppose that $\sin a + \sin b = \sqrt{5/3}$ and $\cos a + \cos b = 1$. What is $\cos(a - b)$?

    (A) $\sqrt{\frac{5}{3}} - 1$ **(B)** $\frac{1}{3}$ **(C)** $\frac{1}{2}$ **(D)** $\frac{2}{3}$ **(E)** 1

sufficiently difficult so that the average score in 2007 was only about three out of 15. The final stage of the AMC series is the USA Math Olympiad (USAMO), a nine-hour proof-based contest taken by 300–500 of the highest scorers on the earlier contests.

An obvious limitation of using AMC scores to assess math achievement is that the test is given to a small subset of students. Approximately 4 million U.S. students per year start high school. About 1.5 million of the 1.8 million who are graduating and going on to college take the SAT. Only about 50,000 high school seniors take the AMC 12. Thus, we will be cognizant of potential selection effects at various points.

The primary advantage of the AMC relative to more standard data sources is that other tests are not designed to distinguish among students at very high performance levels. The math SAT, for example, has limited replicability at the high end. Obtaining an 800 score usually requires making zero mistakes when answering 54 questions at a rate of one minute and 18 seconds per question. Making just three mistakes will drop a student to the 710–750 range. Consequently, it is not surprising that students who get a perfect 800 and then retake the SAT only average 752 on the retake. This is in the 97th percentile, and it is only a little higher than the 741 average achieved by students who retake the math SAT after scoring 760. Hence, we think of the SAT as having limited power for distinguishing students in percentiles above the 97th.

Data on students who take both the AMC 12A and the AMC 12B, in contrast, clearly show that the test is measuring something in a replicable way even at much, much higher percentiles. Students who scored 95 to 105 on the AMC 12A averaged 103 (standard deviation 11) on the AMC 12B. Students who scored 115 to 125 on the AMC 12A averaged 119 (standard deviation 10) on the AMC 12B. And students who scored 138 or higher on the AMC 12A averaged 131 (standard deviation 10) on the AMC 12B. Hence, their average scores on the retake are in the 99.7th percentile of the overall AMC 12 score distribution and probably well above the 99.9th percentile in the whole distribution of U.S. twelfth graders. There has not been any academic work assessing whether the combination of knowledge, problem-solving skills, and test preparation that the AMC measures is predictive of success in college and beyond. We do know, however, that some colleges pay attention to high AMC scores, and the data in our paper Ellison and Swanson (2009) indicate that a very high score on the AMC is an even stronger predictor of future success on the math SAT than is a previous perfect score on the SAT.

Other common standardized math tests are even less useful than the SAT for identifying high-achieving students. On the National Assessment of Educational Progress (NAEP), the primary resource supported by the U.S. Department of Education, even the questions that are classified as "hard" seem straightforward, and Hyde, Lindberg, Linn, Ellis, and Williams (2008) find that state proficiency tests are easier than the NAEP. The two most common tests for international comparisons are the PISA and Trends in International Mathematics and Science Study (TIMSS).[2] PISA is given to 15 year-olds and TIMSS to students in fourth and eighth grades and at the end of high school. Both are administered to representative samples of students in dozens of countries. However, PISA does not appear to be designed to test advanced math skills. TIMMS does have an advanced math test, but it is only administered to students pursuing advanced math courses. The universally administered math component is a "mathematics literacy" test similar to PISA.

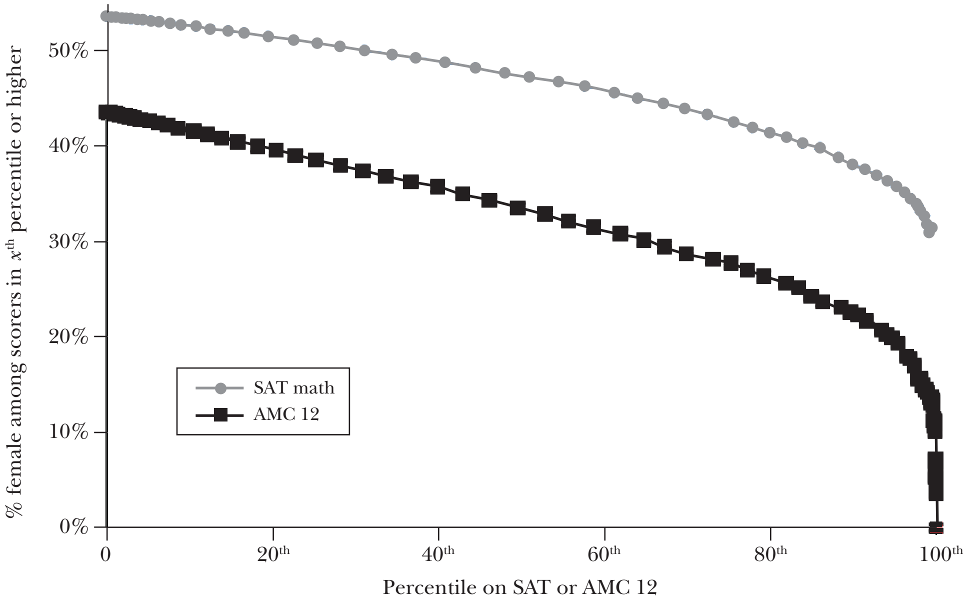
The Gender Gap on the SAT and AMC 12

The descriptive statistics in this section focus on the relative number of girls and boys who reach various levels of performance. Specifically, writing $n_{\mathrm{f}}(\tau)$ for the total number of females with scores of at least τ and $n_{\mathrm{m}}(\tau)$ for the number of males with such scores, the graphs in Figure 3 show the percent female, $100n_{\mathrm{f}}(\tau)/(n_{\mathrm{f}}(\tau) + n_{\mathrm{m}}(\tau))$, as a function of τ.

The top curve is a benchmark derived from data on the math SAT scores of 2007 college-bound seniors. A rescaling of math SAT scores by percentile is on the

[2] The Appendix to this paper, available with the paper at ⟨http://www.e-jep.org⟩, offers more background for comparing the NAEP, PISA, and TIMSS. Figure A1 in the Appendix reproduces the five most difficult of the 21 publicly released "hard" questions from the 2005 NAEP. The set includes some questions on middle school topics that require a little bit of reasoning and a couple questions on topics from standard college-preparatory mathematics that are completely straightforward. Figure A2 in the Appendix contains a sample question from PISA. Figure A3 in the Appendix presents a sample question that appeared on both PISA and TIMSS.

Figure 3
Gender Gap on SAT and AMC 12



Source: Data for the SATs are taken from CollegeBoard.com (2010).
Note: The top curve is derived from data on the math SAT scores of 2007 college-bound seniors. The bottom curve used data on students at American schools taking the 2007 AMC 12. Interested readers should refer to the Appendix to this paper, available with the paper at ⟨http://www.e-jep.org⟩, for a detailed description of the data. When students participated in the AMC 12 (or 10) multiple times, the latter of the two scores is included here and in subsequent analyses.

x-axis, and the percent female among students scoring at each level or higher is on the *y*-axis. The fact that the curve starts out above 50 percent on the left side reflects that more girls take the SAT: the raw numbers are about 800,000 versus 700,000. The fraction female drops to 50 percent around the 30th percentile, reflecting that the number of boys and girls achieving scores in excess of 460 are approximately equal. The percent female drops substantially at higher SAT scores. Approximately 200,000 boys and 150,000 girls receive scores of at least 600. The percent female declines most steeply at the highest percentiles and drops to 31 percent for students at the highest percentile. This reflects that about 2.1 times as many boys as girls score 800.

We constructed the bottom curve of Figure 3 in a similar way, using data on students at American schools taking the 2007 AMC 12. The scaling convention is mechanically identical to that of the SAT curve—the *x*-axis is linear in percentile ranks within the population of U.S. AMC 12 takers. The populations taking the two tests are quite different, however, so readers should keep in mind that the percentiles have very different meanings. Several aspects of the graph are noteworthy.

First, the left-most point of the graph shows that 44 percent of AMC 12 test takers are female. This indicates that, in the aggregate, high-achieving girls and boys are roughly equally likely to participate in the AMC 12. Most AMC takers come from the high end of the SAT population and the population of students with SAT scores of 600 or above is 43 percent female.

Second, the gender gap is larger on the AMC than on the SAT when one looks at comparable performance levels at the high end of the range that the SAT can measure. We find a 4.2 to 1 male–female ratio at the 100 AMC 12 level (100 out of a possible score of 150), which is a substantially larger gender gap than the 2.1 to 1 ratio at the roughly similar 800 SAT level. This suggests that there may be differential selection into AMC taking, with girls of very high achievement being substantially less likely to take the AMC than comparably accomplished boys. An alternate possibility is that among students who know the SAT material equally well, girls may be less likely to have learned the additional material and developed the problem-solving skills needed to achieve a high AMC score.[3] Some such effect must be substantial and we regard it as an important finding. Math contests are one of the main institutions motivating high-ability students to go beyond the standard high school curriculum and develop greater knowledge and problem-solving skills. If math contests are less appealing to girls than to boys, then this will be a reason why fewer girls are reaching very high achievement levels.

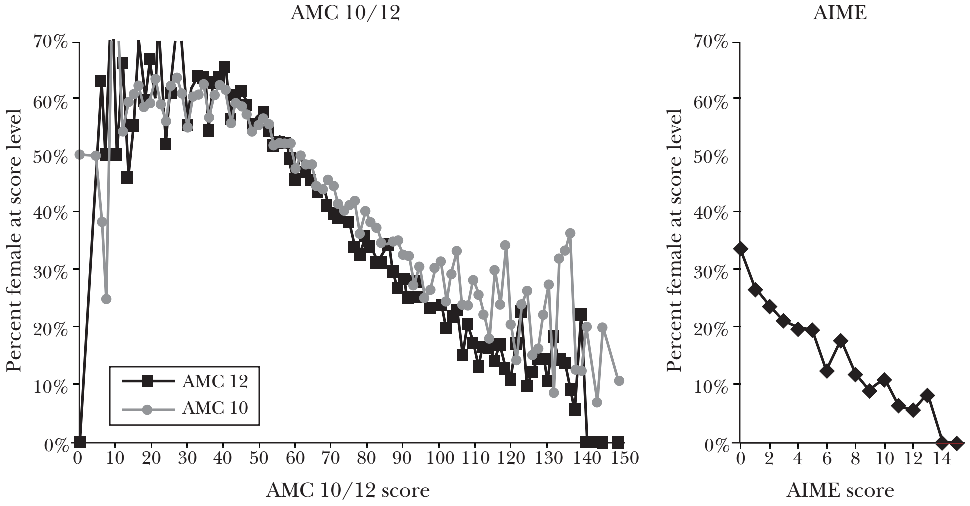
A third observation from Figure 3—indeed, the one that is most visually striking—is that the AMC curve turns sharply downward at very high scores that represent the percentiles above those that the SAT can measure. The percent female falls to 14 percent for students in the 99th percentile of the U.S. AMC 12 population (1,112 students with scores of 114 or higher) and 10 percent (a 9 to 1 male–female ratio) in the 99.9th percentile (a population of 206 students with scores of 130.5 or higher). The top 46 scorers were all male.[4] At these extreme score ranges, it becomes increasingly implausible that gender-related selection into taking the AMC could account for much of the effect: why would girls capable of scoring 130 or higher be only one-quarter as likely to take the test as boys who would do this well? Indeed, the knowledge and problem-solving skills needed to get a 130 are sufficiently high so that we feel that almost all students (male or female) who have acquired such skills are probably taking the AMC 12, making gender-related selection nonexistent (for measurements of what the gender composition of high scorers would be under universal administration).

We interpret the body and right tail of this distribution in combination as suggesting that many very talented girls are simply not taking the AMC 12 and that there is an additional effect in which a smaller fraction of girls who do become involved in math contests develop the mathematical knowledge and problem-solving skills necessary to achieve extremely high scores. We would like to emphasize that

[3] In addition, a portion of the effect may be attributable to the AMC test being more accurate at identifying the 99th percentile of the SAT-taking population (as opposed to the SAT's mixing in many 97th percentile students). A portion could also be attributable to differences between the tests. Hyde, Fennema, and Lamon (1990), for example, discuss evidence suggesting that the gender gap was larger on tests requiring complex problem solving.

[4] In the full dataset only the top 24 scorers were male: one Canadian girl scored 144.

Figure 4
Gender Gaps on AMC 10, AMC 12, and AIME



Note: The left panel of Figure 4 compares the AMC 10 and AMC 12 gender gaps. The right panel shows the gender gap on the AIME.

nothing in this data requires an assumption of different underlying distributions of *ability* in the full male and female populations, as achievement reflects both ability and educational investments.

The Gender Gap on Other AMC Contests

Other AMC contests are a source of complementary information on the gender gap among younger students and very high achievers.

First, the AMC 10 is a slightly easier contest taken by students in grades ten and below. The left panel of Figure 4 compares the AMC 10 and AMC 12 gender gaps. Note that we have changed a couple of things to facilitate comparisons: the *x*-variable in these graphs is the AMC score rather than the percentile; and the *y*-variable is now the fraction female among students at each score level, rather than at the score level or above. Two AMC 12 patterns are easier to see in this graph. First, the population of students receiving just about every score below 58.5 is more than half female. This suggests that the differential selection into AMC taking is such that girls of moderate accomplishment are more likely to take the AMC than are boys who would do about as well. Second, the percent female declines fairly smoothly as we move through the range in which most of the data lie. The AMC 10 data are fairly similar to the AMC 12 data. One notable difference, however, is that the percent female drops off somewhat more slowly at scores above 70 and is substantially higher at the highest score levels. This pattern could be due to differences in the tests. Part of it could also be due to differential selection effects: for example, extreme high-achieving girls may be less likely than extreme high-achieving boys to decide to "compete up" (that is, take the

AMC 12 instead of the AMC 10 in earlier grades). But the most obvious difference between the tests is that the AMC 10 is being taken earlier in high school, which suggests that the effects that lead to the high-end gender gap in high school build throughout the high school years.

The saw-tooth pattern one sees in the high score range of these graphs may also be of interest in connection with the literature on risk taking by boys and girls: the source of the pattern is that the fraction female is higher at scores that can only be achieved by leaving answers blank and lower at scores that are obtained by filling in an answer on every question. For example, one-third of the girls who correctly answered 24 of the 25 questions left the final answer blank, whereas only 13 percent of the boys with 24 correct answers did so.

The right panel of the figure is a similar graph of the gender gap on the AIME. Recall that the AIME is taken only by students who have first achieved a very high score on the AMC 10 or 12, so this can be thought of as an additional look at the gender gap in the right tail. The test emphasizes the ability to solve hard problems over speed; the exam is three hours long and the median participant in 2007 only solved three of the 15 problems. The pool qualifying to take the AIME is 22 percent female. The graph illustrates that the percentage female declines smoothly as one looks at higher score levels. The right-tail results are similar to what one finds on the AMC 12. There were just five girls versus 80 boys scoring eleven or above on the AIME, which is very similar to the five girls versus 90 boys scoring 136.5 or higher on the AMC 12.

Cross-Sectional Patterns in the Gender Gap within the U.S.

Several recent papers have examined the gender gap in multicountry datasets. The magnitude of the gap has been found to vary from country to country, which may reflect both general cultural differences and differences in educational institutions (Andreescu, Gallian, Kane, and Mertz, 2008; Guiso, Monte, Sapienza, and Zingales, 2008; Machin and Pekkarinen, 2008; Penner, 2008). In this section, we explore a related topic—how the high-end gender gap varies from school to school within the United States. We provide statistically significant evidence that the gender gap is narrower in some schools than others, but find that the magnitude of the variation is not very large.

The raw data make it immediately apparent that the existence of some gender gap is almost universal (at least among high-performing schools). For example, 126 schools had eight or more students score above 100 on the AMC 12. At 122 of these 126 schools, boys outnumbered girls among the high scorers.[5] It is important to keep in mind that the 4 percent of high-achieving schools that had no gender

[5] The exceptions are one private school and three very strong but otherwise unremarkable public schools: at Holmdel High School (Holmdel, New Jersey), eight of the 16 high scorers were girls; at Canton High School (Canton, Massachussetts), five of the nine high scorers were girls; and at Lawton Chiles High School (Tallahassee, Florida), five of the nine high scorers were girls. At the private Hotchkiss School (Lakeville, Connecticut), six of the eleven high scorers were girls.

gap in realized achievement does not provide a valid estimate of the number of schools without a gender gap in *expected* performance: in a Poisson-like model, many more schools than this would randomly have more girls than boys among their high scorers even if the boys were twice as likely to succeed.

Our first formal analysis is like the simple "dartboard" analysis of geographic concentration in Ellison and Glaeser (1997): we ask whether there is any evidence of variation in the gender gap after one takes out the variation that would arise purely at random. Specifically, the way we formalize "at random" is to suppose that the environment of school i is such that the number of high-scoring girls f_i will have a binomial distribution (N_i, p_i), where N_i is the total number of high-scoring students at the school and p_i is a parameter that reflects how the environment affects the gender gap (p_i is the probability that a high-scoring student from school i will be female). The purely random benchmark would be that there is no variation in p_i across schools. The alternative is that p_i does vary across schools—it is larger in some schools that have successfully created an environment that leads to a smaller gender gap, and it is smaller in others. To provide a formal test of the purely random model and (in the alternative) to estimate the degree to which p_i does appear to vary across schools, we assume that the p_i are themselves independent realizations from a Beta (α, β) distribution and estimate the parameters of this distribution by maximum likelihood.

We estimated the mean and variance of the p_i in this manner using data on the number of girls and boys in each school scoring above 100 on the AMC 12.[6] The point estimates are that the p_i are drawn from a distribution with a mean of 0.18 and a standard deviation of just 0.05. Recall that the p_i can be interpreted as the probability that a high-scoring student from school i will be female. The fact that the standard deviation is positive and statistically significant implies that we can reject the purely random model: there are some schools where p_i is bigger (the gender gap is smaller) and others where p_i is smaller (and the gender gap is bigger). But from a practical perspective, the more important thing to take away is that the estimated standard deviation indicates that there is only moderate variation in the gender gap from school to school: there can't be many schools that don't have a substantial gender gap. We also carried out a similar calculation using data on the number of boys and girls scoring above 120 on the AMC 12. Here, the probability that a high scorer is a girl is estimated to be 0.11 in the mean school with a standard deviation 0.03 across schools. This test does not provide statistically significant evidence that the gender gap varies across schools, but it would not be expected to have much power because the number of girls scoring above 120 on the AMC 12 is so small—in fact, only eight U.S. schools have more than one girl scoring above 120 on the AMC 12.

We conclude that the factors that are contributing to the gender gap are felt quite broadly. The gender gap is bigger at some schools and smaller at others, but

[6] In addition to schools with no students scoring above 100, we drop schools outside the United States, schools we were not able to identify in the NCES data (as explained in the Appendix at ⟨http://www.e-jep.org⟩), and single-sex schools, leaving a sample size of 1,307 schools.

there are only moderate differences across schools and it appears that almost all high schools have a substantial gender gap.

While the variation in the gender gap across schools is small as a fraction of the gender gap, it is substantial when one thinks about it in relation to the number of girls who are currently achieving high scores. A school that produces 0.75 high-scoring girls and five high-scoring boys per year would be 13 percent female on average among high scorers, whereas 1.5 girls and five boys would be 23 percent female. Hence, if some set of policy changes could shift a school from being one standard deviation below average to being one standard deviation above average, it would roughly double the number of high-scoring girls at that school. From this perspective, it seems important to investigate where the gender gap is relatively large and where it is relatively small. Our random model, however, points out that this will be difficult: most of the variation across schools will be pure random noise, and it may be hard to find systematic patterns.

When we examined the determinants of the gender gap using simple school-level regressions, we failed to find many statistically significant patterns. For each public school that could be matched to NCES data, we computed the fraction female among students in the school who scored at least 100 (and 120) on the AMC 12. As our explanatory variables, we used a number of characteristics of the school and of the zip code such as the race and ethnicity of the school population and the education and income of the student's zip code. Almost none of the estimates were significant at the 5 percent level.

One interesting estimate is that the gender gap appears to be somewhat narrower in schools that have many high achievers on the AMC 12. In fact, when this variable is included on the right-hand side of the regression in log form, it is statistically significant at the 5 percent level. Table 1 looks at this last relationship more closely by dividing schools into bins on the basis of the number of students scoring at least 100 on the AMC 12 and tabulating the number of high-scoring boys and girls within schools in each bin. The first set of columns show that the percent female among students scoring at least 100 on the AMC 12 rises from about 15 percent in the lowest bin to over 20 percent in most of the bins containing high-performing schools. The second set of columns tabulate the number of boys and girls scoring at least 120 on the AMC 12. Here, the percent female is highest in the schools just below the two best, but the pattern is not as clear. The third column examines students scoring at least 130 on the AMC 12. It contains the most striking pattern: there are no girls at all (versus 49 boys) in the lowest four bins and eight girls (versus 23 boys) in the top four.

Evidence from the Extremes: U.S. and International Comparisons

In this section, we examine extreme high-performing students who are chosen to represent their countries in international competition. The motivation for this section is both that one can make cross-country comparisons in this way and also that looking at the extremes may help us understand the gender gap at less extreme

Table 1

Patterns in the Gender Gap among High Scorers on the AMC 12 across Schools

		Gender composition of high-scoring students in schools within bin								
		AMC 12 $\geq$ 100			*AMC 12 $\geq$ 120*			*AMC 12 $\geq$ 130*		
Number of students from school with AMC 12 $\geq$ 100	*Number of schools*	*Female*	*Male*	*% Female*	*Female*	*Male*	*% Female*	*Female*	*Male*	*% Female*
1	434	65	369	15.0	2	28	6.7	0	6	0
2	187	71	303	19.0	1	36	2.7	0	15	0
3–4	159	88	449	16.4	4	39	9.3	0	13	0
5–6	62	68	272	20.0	6	27	18.2	0	15	0
7–10	64	103	410	20.1	6	56	9.7	3	19	13.6
11–15	28	74	272	21.4	3	36	7.7	1	15	6.3
16–25	13	61	173	26.1	6	25	19.4	2	9	18.2
26–40	5	35	122	22.3	9	23	28.1	5	8	38.5
41–60	2	22	64	25.6	4	16	20.0	3	8	27.3
61–100	1	14	84	14.3	1	14	6.7	0	4	0
>100	1	25	95	20.8	1	10	9.1	0	3	0

percentiles. Indeed, the U.S. data reveal one striking contrast that may turn out to be an important observation.

U.S. Teams

The AMC series includes one additional invitational test beyond the AMC 12 and AIME: the United States of America Mathematical Olympiad (USAMO). The rules for invitations have varied from year to year, but usually, 300 to 500 of the students with the highest AIME scores take the USAMO. The USAMO is very different from the AMC and AIME: whereas the AMC and AIME exams mostly require knowledge of material from the standard college-preparatory curriculum, the USAMO is a proof-oriented contest and therefore relies on skills that would not be obtained from high school coursework, except perhaps at a handful of U.S. high schools.

Since 1974, the United States has sent teams to compete in the International Mathematics Olympiad. Over the full 35-year period, the United States has sent 224 students to the IMO. Five have been female (actually there were just three female students, two of whom went twice). In the first 24 years there were no female team members. Since then the ratio has been 12 to 1 male to female.

Since 2007, the U.S. has also sent an eight-person team to the China Girls' Math Olympiad. These teams are publicly announced, which provides us with an opportunity to compare the backgrounds of a group of extreme high-scoring girls with that of a group of extreme high-scoring boys: the CGMO team members are roughly the top-scoring girls from the USAMO, whereas the IMO team is roughly the top-scoring students regardless of gender.[7] We do not have data on individual USAMO scores, but the highest-scoring CGMO team member, Sherry Gong in 2007, was also on the IMO team and hence must have been in the top twelve. No other CGMO student, however, was in the top 24, and announced cutoffs suggest that the lowest-scoring 2008 CGMO team member was approximately 170th on the USAMO. The CGMO team members are somewhat closer to the IMO team members on the other AMC tests. For example, in 2007 the median CGMO team member scored nine on the AIME and the median IMO team member scored ten.

Table 2 presents data on the schools which produced IMO and CGMO members in 2007 and 2008. Specifically, it reports the number of each student's classmates who scored at least 100 on the 2007 AMC 12, the number who scored at least five on the 2007 AIME, and the number who qualified to take the 2008 USAMO along with percentile ranks of the schools on these measures.

The bottom half of the table contains statistics on the schools of team members for the China Girls Math Olympiad. The nonrepresentativeness of the schools these girls come from is startling: the median CGMO team member comes from a school at the 99.3rd percentile among AMC participating schools, that is, from one of the top 20 or so schools in the country. Only three come from schools that are not in the 99th percentile in most measures. And even those three are from schools that had at least one other student qualify for the 2008 USAMO and are at least in the 91st percentile in terms of the number of high-scorers on the AMC 12.

The male team members for the International Math Olympiad, in contrast, come from a much broader set of schools. Some are from super-elite schools and most come from schools that do very well on the AMC 12, but the median student is just from a 93rd percentile school. The majority of the IMO team members had no schoolmates qualify to take the USAMO, whereas all CGMO team members had at least one schoolmate qualify and most had at least four.

The fact that the top boys and girls are coming from such different sets of schools suggests that one reason why the gender gap is so wide at the highest achievement levels is that the boys are effectively being drawn from a much larger pool. It may be that parents of extremely talented girls are much more likely than parents of extremely talented boys to send them to schools with elite math programs. But we feel that it is implausible that there are not many more highly talented girls in the 99 percent of schools that are not in the top one percent of schools who could also have reached performance levels similar to those of the CGMO team members with the right encouragement and education.

A quick look at the names of the CGMO team members indicates that they are also drawn from a small subset of the population in another dimension: almost all are Asian-American. Andreescu, Gallian, Kane, and Mertz (2008) note that most U.S. IMO team members are also selected from a small fraction of the population

[7] Neither description is exactly right. The IMO team is chosen from among the twelve high scorers on the USAMO using USAMO scores and another test. The first CGMO teams were based on scores in the previous year, and at least one student offered a place on the CGMO team declined.

Table 2
The 2007 and 2008 U.S. IMO and CGMO Teams: Statistics on Team Members' Schools

		School strength: counts and percentile					
Student	*High School*	*Classmates with AMC 12 ≥100*	*School %ile (AMC)*	*Classmates with AIME ≥ 5*	*School %ile (AIME)*	*Classmates with USAMO ≥ 0*	*School %ile (USAMO)*
U.S. International Math Olympiad Teams: 2007 and 2008							
Sherry Gong	Phillips Exeter Acad.	40	99.8	24	99.9	16	100
Eric Larson	South Eugene HS	7	94.9	1	81–93	0	0–91
Brian Lawrence	Mongomery Blair HS	42	99.9	19	99.8	9	99.9
Tedrick Leung	North Hollywood HS	10	97.3	2	93–97	0	0–91
Arnav Tripathy	East Chapel Hill HS	5	91.7	1	81–93	1	91–98
Alex Zhai	University Laboratory HS	5	91.7	0	0–81	0	0–91
Paul Christiano	The Harker School	23	99.5	13	99.7	4	99.2
Shaunak Kishore	Unionville-Chaddsford HS	0	0	0	0–81	0	0–91
Evan O'Dorney	Venture (Indep. Study)	0	0	0	0–81	0	0–91
Colin Sandon	Essex HS	3	83.8	2	93–97	0	0–91
Krishanu Sankar	Horace Mann HS	9	96.6	2	93–97	0	0–91
Alex Zhai	University Laboratory HS	5	91.7	0	0–81	0	0–91
U.S. China Girls Math Olympiad Teams: 2007 and 2008							
Sway Chen	Lexington HS	17	99.1	8	99.4	4	99.2
Sherry Gong	Phillips Exeter Acad.	40	99.8	24	99.9	16	100
Wendy Hou	Hillsborough HS	5	91.7	0	0–81	1	91–98
Jennifer Iglesias	IL Math & Sci. Acad.	45	99.9	12	99.6	4	99.2
Colleen Lee	Palo Alto HS	27	99.6	12	99.6	6	99.7
Patricia Li	Lynbrook HS	44	99.9	13	99.7	6	99.7
Marianna Mao	Mission San Jose HS	18	99.3	10	99.5	4	99.2
Wendy Mu	Saratoga HS	10	97.3	6	99.1	7	99.8
In Young Cho	Phillips Exeter Acad.	40	99.8	24	99.9	16	100
Jenny Jin	The Taft School	7	94.9	5	98.7	2	98–99
Carolyn Kim	Lawton Chiles HS	8	95.8	2	93–97	1	91–98
Jennifer Iglesias	IL Math & Sci. Acad.	45	99.9	12	99.6	4	99.2
Colleen Lee	Palo Alto HS	27	99.6	12	99.6	6	99.7
Wendy Mu	Saratoga HS	10	97.3	6	99.1	7	99.8
Lynnelle Ye	Palo Alto HS	27	99.6	12	99.6	6	99.7
Joy Zheng	Lakeside School	18	99.3	10	99.5	4	99.2
Median for male IMO team members		*6*	*93.3*	*1.5*	*81–93*	*0*	*0–91*
Median for CGMO team members		*18*	*99.3*	*10*	*99.5*	*4*	*99.2*

Note: Table 2 presents data on the schools which produced IMO and CGMO members in 2007 and 2008. Specifically, it reports the number of each student's classmates who scored at least 100 on the 2007 AMC 12, the number who scored at least five on the 2007 AIME, and the number who qualified to take the 2008 USAMO along with percentile ranks of the schools on these measures. The percentile rank is always the rank that the school would have without the student in question. We do this because otherwise all schools would have a very high rank on the USAMO qualifier metric. We use 2007 data rather than the most recent data for the AMC and AIME because we need to compute school-level percentiles using our complete dataset and this only runs through 2007.

in that many are Asian, Jewish, children of immigrants, and/or children of parents with advanced mathematical training. The fact that CGMO team members are mostly Asian-American may or may not involve a differential selection effect contributing to the gender gap. However, it is a further observation suggesting that there is a substantial pool of potentially talented mathematicians who are not being brought up to the highest level (both in and out of elite high schools).

International Evidence

The International Math Olympiad was first held in 1959. Over the past 50 years it has grown from a small contest among Soviet-bloc nations to a worldwide contest among 100 countries. Each country may send up to six high school students. These students are often winners of the country's national contest, although the manner in which teams are selected varies. Andreescu, Gallian, Kane, and Mertz (2008) analyze the gender composition of IMO teams in order to gain insight into the degree to which the gender gap is due to cultural, educational, and other factors that vary across countries. They show that there are statistically significant differences in the gender gap across countries and emphasize the outliers in their discussion. They argue: "Girls were found to be 12%–24% of the children identified as having profound mathematical ability when raised under some conditions; under others, they were 30-fold or more underrepresented. Thus, we conclude that girls with exceptional mathematical talent exist; their identification and nurturing should be substantially improved so this pool of exceptional talent is not wasted."

We see our CGMO results as in complete agreement with their view that the U.S. educational system is failing to develop a substantial pool of girls who possess exceptional mathematical talent. But we see the broad patterns of the IMO data differently. Where Andreescu et al. (2008) emphasize the statistical significance of differences across countries, we would emphasize that the magnitudes of the differences across countries are small.

Our primary IMO data are the data of Andreescu, Gallian, Kane, and Mertz (2008), in particular, the gender of each student who participated at some point in 1988–2008 as a member of the team from one of 30 high-scoring countries.[8] One basic fact about the IMO is that there is a very large gender gap: only 5.7 percent of the participants in this sample (185 of 3,246) are female. There has been some narrowing of the gender gap over time: the fraction female increases from 4.3 percent in 1988–97 to 6.8 percent in 1998–2008. For comparison, the 6.8 percent figure is roughly similar to the gender gap for U.S. students at the 135 and above level on the 2007 AMC 12 and at the 11 and above level on the 2007 AIME—scores that were achieved by 117 and 54 U.S. students, respectively.

Andreescu et al. (2008) highlight the outliers in the data—the team from Yugoslavia/Serbia and Montenegro is 24 percent female in the most recent decade whereas Iran, Japan, and Poland sent entirely male teams. Looking at the magnitudes of the differences, however, we would emphasize that the gender gap is, if anything, strikingly universal. In the 27 countries they consider for the 1998–2008

[8] We also analyze data on additional countries collected by Matjaz Zeljko, which is posted on the IMO website: ⟨http://www.imo-official.org⟩. Our sample has 26 countries rather than 30 because we combine Germany/East Germany/West Germany and Czechoslovakia/Czech Republic/Slovakia into single countries.

period, the total number of female participants per team had a mean of 4.6 (out of 66 total participants) and a standard deviation of 3.7. In a model in which each participant was female with independent probability 0.069, the number of female participants from a country that sent 66 students would have mean 4.5 and standard deviation 2.1. Some heterogeneity across countries will be needed to account for the 17 female participants from Yugoslavia/Serbia and Montenegro and the countries sending no young women, but the magnitude cannot be very large.[9] Whatever combination of factors is leading to the gender gap at the extreme, it appears to be widespread. (As before, the magnitude of the variation across countries is small relative to the size of the gender gap, but not small relative to the number of women who are currently reaching the IMO.)

Several studies have attempted to test whether the gender gap in various countries can be related to measures of cultural, political, and economic gender equity. For example, Guiso, Monte, Sapienza, and Zingales (2008), using PISA data, find that the gender gap in average scores is smaller in countries with greater gender equity. We looked for a similar effect in the IMO data by regressing the number of female IMO competitors in 2006–2008 on the World Economic Forum's Gender Gap Index for 91 countries. However, the regression has little explanatory power and the positive point estimate on the gender gap index is not statistically significant. We should note, however, that Hyde and Mertz (2009) independently conducted a similar test and found a positive significant correlation. The different finding appears to stem from two differences in the samples: they restrict attention to the highest-performing countries and use more years of data. The longer time horizon increases the precision of the estimates. And the restriction to high-performing countries eliminates some countries that rank very highly on gender equity index but have unexceptional female IMO participation.

Conclusion

The American Mathematics Competitions are able to draw consistent distinctions between the problem solving and precalculus math skills of students even at very high percentiles of achievement and hence provide an opportunity to learn more about what goes on in the upper part of the distribution. With this data and results of other mathematics competitions, we verify the common observation that there is a large gender gap among high math achievers. We are also able to enrich existing descriptions on several dimensions. We have noted that the gender gap widens substantially at percentiles above those that conventional tests can measure

[9] That is, the random model is close enough to what we see that there can't be too much additional variation across countries. To provide a formal estimate, we perform the same calculation as earlier. We assume that the probability that each participant from country i is female is p_i and estimate the mean and standard deviation of the p_i under the assumption that these have a Beta distribution across countries. In the 1998–2008 time period, we estimate the mean of p_i to be 0.065 and the standard deviation of p_i to be 0.031. The standard errors on these estimates are 0.009 and 0.011, respectively.

well, that the gender gap seems widespread both across U.S. schools and in international comparisons, and that high-achieving girls are concentrated in schools with elite math programs (which comes out both in the CGMO data and in the data on students scoring 130 on the AMC 12).

We have consciously focused on reporting the facts in our data rather than on attempting to draw out what the data might say about the relative importance of the many different factors that may contribute to the gender gap. Mostly we do this because our data contain many new facts and do not seem particularly well-suited for distinguishing theories. Indeed, our evidence offers some warnings about certain kinds of speculation. For example, we believe there is limited value to putting a lot of effort into using math test scores for estimating "ability" distributions, especially at the highest achievement levels, when almost all girls who would be capable of achieving extremely high scores do not do so.

However, when pushed to speculate on underlying causes, it does seem that several elements in our results are consistent with the view that girls suffer in becoming high achievers in mathematics because they are more compliant with authority figures and/or are more sensitive to social environment. In most high schools, even in the highest-level "honors" courses, it is probably unusual to teach material at the level needed to bring students to the 99th percentile. If talented girls are less likely to complain and get schools to make special accommodations, and if social factors make them less likely to join math teams or take advanced online courses, then they will be more underrepresented when we examine achievement levels that are further beyond those developed in the standard classroom setting. Such explanations could also fit well with our observations on the China Girls' Math Olympiad participants: schools that have very large numbers of high-achieving students can be places where girls can join a community learning advanced material.

A number of alternative explanations for the gender gap are also possible. One would be a model in which there is less variance of ability in the female population. This could be consistent both with the increasing gender gap at the highest achievement levels and with the extreme high-achieving girls coming from the extreme high-achieving schools. For example, consider a simple additive model in which achievement is the sum of ability and educational quality, with educational quality and ability each being normally distributed, and with the variance of ability being greater in the male population (and the variance of education being gender independent). In such a model the male–female ratio would increase as one looked further out in the right tail of achievement. And conditional on a high level of achievement, the expected educational quality would be higher for females, which could be a way to explain our finding that high-scoring girls are concentrated in high-achieving schools. Another alternate model would be a model in which a lack of girls in the population of extreme high math achievers is not a bad thing: it might be that the girls who could reach the highest achievement levels tend not to do so because they are more likely to have other skills and interests as well and they tend to pursue less math-focused paths that lead them to develop portfolios of skills that will be more valuable in the long run. This could be part of what is going on, although we would note that the math skills needed to score 100 on the AMC 12 do not look very high in comparison, for example, to what is needed to succeed in the economics profession.

We would also point out that it would be desirable for theories to explain math achievement and other gender comparison facts at the same time. For example, the male–female ratio among students scoring 800 on the SAT Critical Reading test is about 1 to 1, and the ratio on the SAT Writing test is about 0.7 to 1. With regard to the compliance/community explanation, it could be that the verbal SAT tests do not test much beyond what is gained from a standard high school English class, plus a lot of reading, and hence compliance with standard school path is not costly. But further research on a range of topics seems worthwhile before attempting to draw conclusions from few facts.

Our observations that gender gaps in high math achievement appear to be present in almost every school and country may be seen as discouraging. One should keep in mind, however, that even the modest variation we've found across schools and countries is sufficiently large so that it may not be too difficult to make large proportionate increases in the number of girls who are doing well in math. Our comparison of schools with different achievement levels is particularly hopeful in that it suggests that one might be able to increase the number of students reaching high achievement levels and simultaneously narrow the gender gap by increasing the number of schools that provide opportunities for elite math achievement. Further studies of the environments where girls (and boys) are doing relatively poorly and relatively well may be very valuable.

■ *This project would not have been possible without Professor Steve Dunbar and Marsha Conley at AMC, who provided access to the data as well as their insight. We also thank Mary O'Keeffe, Janet Mertz, and the JEP editors for extensive comments, David Card and Jesse Rothstein for help with data matching, and Syndie Kim for research assistance. Financial support was provided by the NSF (SES-05508897), the Toulouse Network for Information Technology, and Microsoft Research.*

References

Andreescu, Titu, Joseph A. Gallian, Jonathan M. Kane, and Janet E. Mertz. 2008. "Cross-Cultural Analysis of Students with Exceptional Talent in Mathematical Problem Solving." *Notices of the American Mathematical Society*, 55(10): 1248–60.

Blau, Francine D., and Lawrence M. Kahn. 2000. "Gender Differences in Pay." *Journal of Economic Perspectives*, 14(4): 75–99.

Brown, Charles, and Mary Corcoran. 1997. "Sex-Based Differences in School Content and the Male–Female Wage Gap." *Journal of Labor Economics*, 15(3): 431–65.

CollegeBoard.com. 2010. "SAT Percentile Rank or Males, Females, and Total Group, 2007 College-Bound Seniors—Mathematics." A table. http://www.collegeboard.com/prod_downloads/highered/ra/sat/SAT_percentile_ranks_males_females_total_group_mathematics.pdf.

Ellison, Glenn, and Edward L. Glaeser. 1997. "Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach." *Journal of Political Economy*, 105(5): 889–927.

Ellison, Glenn, and Ashley Swanson. 2009. "The Gender Gap in Secondary School Mathematics at High Achievement Levels: Evidence from the American Mathematics Competitions." NBER Working Paper 15238.

Freeman, Catherine E. 2005: *Trends in Educational Equity of Girls & Women: 2004.* (NCES 2005–016). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U. S. Government Printing Office.

Fryer, Roland G., Jr., and Steven Levitt. Forthcoming. "An Empirical Analysis of the Gender Gap in Mathematics." *American Economic Journal: Applied Economics.*

Guiso, Luigi, Ferdinando Monte, Paola Sapienza, and Luigi Zingales. 2008. "Culture, Gender, and Math." *Science*, 320(5880): 1164–65.

Hyde, Janet S., Elizabeth Fennema, and Susan J. Lamon. 1990. "Gender Difference in Mathematical Performance: A Meta Analysis." *Psychological Bulletin*, 107(2): 139–55.

Hyde, Janet S., Sara M. Lindberg, Marcia C. Linn, Amy B. Ellis, and Caroline C. Williams. 2008. "Gender Similarities Characterize Math Performance." *Science*, 321(5888): 494.

Hyde, Janet S., and Janet E. Mertz. 2009. "Gender, Culture, and Mathematics Performance." *Proceedings of the National Academy of Science*, 106(22): 8801–07.

Machin, Stephen, and Tuomas Pekkarinen. 2008. "Global Sex Differences in Test Score Variability." *Science*, 322(5906): 1331–32.

The Mathematical Association of America, American Mathematics Competitions. 2007. *58th Annual Summary of High School Results and Awards.* http://www.unl.edu/amc/d-publication/d1-pubarchive/2006-7pub/2007-1012Summary.pdf.

Organization for Economic Co-operation and Development (OECD). 2006. *PISA 2006: Science Competencies for Tomorrow's World.* Paris: OECD.

Penner, Andrew M. 2008. "Gender Differences in Extreme Mathematical Achievement: An International Perspective on Biological and Social Factors." *American Journal of Sociology*, 114(Suppl.): S138-S170.

Perie, Marianne, Rebecca Moran, and Anthony D. Lutkus. 2005. *NAEP 2004 Trends in Academic Progress: Three Decades of Student Performance in Reading and Mathematics* (NCES 2005-464). U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Washington DC: Government Printing Office.

U.S. Department of Commerce, Bureau of the Census. 2000. *Census of Population and Housing, 2000: Summary File 1.* Washington, DC.

U.S. Department of Commerce, Bureau of the Census. 2000. *Census of Population and Housing, 2000: Summary File 3.* Washington, DC.

U.S. Department of Education, National Center for Education Statistics. *Common Core of Data: Public School Data, 2006–2007.* Available at http://nces.ed.gov/ccd/bat/.

U.S. Department of Education, National Center for Education Statistics. *PSS Private School Universe Survey Data, 2006–2007.* Available at http://nces.ed.gov/pubsearch/getpubcats.asp?sid=002.

Xie, Yu, and Kimberlee A. Shauman. 2003. *Women in Science: Career Processes and Outcomes.* Cambridge, MA: Harvard University Press.

This article has been cited by:

1. K. Safarzynska. 2011. Socio-economic Determinants of Demand for Private Tutoring. *European Sociological Review* . [CrossRef]