CESifo
AREA CONFERENCES 2019

Economics of Education

Munich, 30–31 August 2019

Dynamics of the Gender Gap in High Math Achievement

Glenn Ellison and Ashley Swanson

20 YEARS CESifo

Dynamics of the Gender Gap in High Math Achievement[1]

Glenn Ellison
MIT and NBER

and

Ashley Swanson
University of Pennsylvania and NBER

April 2019

[1]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. Daniel Ehrlich provided excellent research assistance. Financial support was provided by the Sloan Foundation and the Toulouse Network for Information Technology.

Abstract

This paper examines the dynamics of the gender gap over the high school years among high scorers on the American Mathematics Competitions. A clear gender gap is already present by 9th grade and the gender gap widens over the high school years. High-achieving students must substantially improve their performance from year to year to maintain their within-cohort rank, but there is nonetheless a great deal of persistence in the rankings. Several gender-related differences in the dynamics contribute to the widening of the gender gap, including differences in the rates with which male and female students stop participating in the contests, and in the mean and variance of year-to-year improvements among those who continue. A decomposition indicates that the most important difference is that fewer girls make large enough gains to move up substantially in the rankings. An analysis of students on the margin of qualifying for a prestigious second stage exam provides evidence of a discouragement effect: some react to falling just short by dropping out of participating in future years, and this reaction may be more common among girls.

Keywords: gender gap, mathematics education, AMC, American Mathematics Competition, educational mobility

JEL Codes: I20, J16

1 Introduction

The gender gap in average science and math achievement by the end of high school has narrowed significantly in recent decades and is qualitatively small today.[1] However, girls are underrepresented among high-achieving math students in middle and high school and this may contribute to their underrepresentation in STEM fields (science, technology, engineering and mathematics), both in college majors and the workforce.[2] These gaps have been shown to vary with potentially manipulable environmental factors such as local culture and the presence of same-gender instructors.[3] To the extent that there is a role for policy in addressing female underrepresentation in STEM, several natural questions arise: at what point in students' development do these gaps occur, how do they evolve over time, and why?[4]

This paper takes advantage of a new panel dataset on American Mathematics Competition (AMC) participants to examine the dynamics of the gender gap over the high school years within this large population of very high-achieving US math students. We begin with two new observations on the gender gap in this population: girls are already quite underrepresented among high-achievers at the beginning of high school; and the gender gap continues to widen over the high school years. We then further explore the latter trend, taking advantage of the panel nature of our dataset to compare the future experiences of boys and girls who performed identically in 9th (and later) grades. We find that several factors contribute to the widening gender gap, with one important message being that the widening gap is not just (or even primarily) about what happens to boys and girls who are already at the top. The largest contributing factor reflects that, among students who are not quite in the top rank groups, girls are less likely than boys to make the large year-to-year improvements needed to move up substantially in the rankings. Finally, we exploit a natural experiment to provide causal evidence on the role of boys' and girls' reactions to disappointment.

The primary data used in this paper are scores on a math contest. In previous work, we argued that two features of the AMC tests distinguish them from other available instruments

[1] See Xie and Shauman (2003) and Goldin et al. (2006) among others.

[2] Joensen and Nielsen (2016) find that an exogenous decrease in the costs of acquiring advanced high school mathematics leads to higher average earnings for girls, with no corresponding effects for boys. See also Hedges and Nowell (1995), Guiso et al. (2008), Hyde et al. (2008), and Ellison and Swanson (2010) on math test scores, and Ginther and Kahn (2004) and Carrell et al. (2010) on workforce issues.

[3] See Guiso et al. (2008), Pope and Sydnor (2010), and Carrell et al. (2010)

[4] As discussed in Fryer and Levitt (2010), boys and girls perform equally well upon entry to school, but over the first six years of school, girls lose more than two-tenths of a standard deviation.

and give them the potential provide more general insight into high math achievement among US high school students. First, the AMC tests are much better than commonly studied tests at identifying and distinguishing among very high-achieving students. Second, the tests are taken by many of the very best math students in the US. The panel dataset used for the first time in this paper also allows us to exploit a third attractive feature: we can get insights on the development of math achievement by seeing how students perform on tests of similarly high difficulty in 9th, 10th, 11th, and 12th grades.

The attractive features of the AMC data come bundled with an important limitation: as Niederle and Vesterlund (2010) emphasize, our study population consists of students who have chosen to participate in a competition. The students may differ from other high achievers in various ways; e.g., they are presumably more interested in competition, and we only see them perform in a competitive environment.[5] As in our previous work, this is relevant to interpreting the finding of this paper that the high achievement gender gap is already large by 9th grade. We are limited in how much we can say about how the size of the gap would change if the test were universally administered, or administered in an environment that was not described to students as a contest. We hope, however, that the limitation is less relevant to the dynamic analysis that we mostly focus on in this paper. The estimates of the dynamics compare subsequent outcomes for boys and girls who have in common that they selected into competing and achieved identical scores in year t. When those same students participate in year $t+1$, we can hope that selection effects and reactions to competition are not very different across genders. In other analyses, we examine gender-related differences in the rates at which such matched students drop out of participating in the AMC contests. That such dropout does occur can be seen as a fourth potential advantage of the AMC environment: it lets us examine dropouts from and new entry into real-world competition by high-achieving boys and girls, potentially complementing the striking observations about the competitive attitudes of ordinary men and women in the lab by Niederle and Vesterlund (2007), Niederle et al. (2013), Buser and Yuan (2018), and others.

Section 2 begins with some more information on the AMC contests and summary statistics on the dataset. It then presents the two most basic observations that motivate the rest of the paper. One is that that there is already a substantial gender gap among high-achieving 9th graders. The second is that the gender gap widens substantially over the high school years. The first observation motivates examining the individual-level persistence in high

[5]See Gneezy et al. (2003) and Iriberri and ReyBiel (2018), among others, on gender differences in performance in competitive environments.

scores: if it is substantial (which we will find), then the 9th grade gap is a large contributor to the end-of-high-school gap and merits further study. The second observation motivates a richer examination of the dynamics of achievement among high-achieving high school students and the gender-related differences in these dynamics. Many potential explanations have been discussed to account for the single fact that boys outnumber girls among high math achievers. A fuller understanding of the dynamics can provide a much larger set of facts that proposed explanations for the large gender gap at the end of high school would need to explain.

Section 3 takes a step back from the focus on gender to provide some initial observations on the dynamics of high achievement in high school. We present both raw transition matrices and regression analyses looking at how students at a given performance level in year t will perform in year $t+1$. We also examine the rates at which high-achieving students appear to drop out of participation in the following year, and the rates at which new "entrants" start participating and perform well. Several observations are important to understanding the environment in which high-achieving high school students are investing in their math skills. One is that performance is highly persistent even when we make what might seem to be very fine distinctions in initial performance. This suggests that there is substantial heterogeneity in performance among students who are often lumped together as a top-coded group. Another is that high-achieving students are substantially improving their mastery of the precalculus mathematics and problem solving skills tested by the AMC contests over the high school years. The magnitude of the average increases and the fact that there are many more lower-ranked students looking to move up than students close to the top implies that high-achieving students must improve substantially to maintain their position, and the probability of making substantial gains relative to one's cohort is low. Making such an improvement presumably requires substantial effort, which suggests that factors that lead to gender-related differences in effort allocation could play a big role in the dynamics of the gender gap.

Section 4 then explores gender-related differences in the dynamics to identify factors that lead to the widening gender gap. Studies of other environments have identified several gender-related differences that could affect patterns of entry, exit, and improvement on the AMC tests. Azmat and Petrongolo (2014) present a helpful review of experimental evidence on these factors: returns to negotiation, preferences for risk and competition, and sensitivity to social cues. We expect preferences for competition to be particularly relevant here: e.g., Niederle and Vesterlund (2007) show that, in a laboratory experiment in which men and

women have equal ability, men are more likely than women to opt for tournament-based compensation.[6] And as noted above, high achievement in the AMC requires a substantial ongoing investment of time and effort, and girls and boys may allocate their time or effort differently across extracurricular activities over the course of high school.[7] Our analyses indeed uncover several distinct differences by gender that any proposed explanation for the gender gap would have to explain. High-achieving girls improve by less on average from year to year than do boys with similar performance levels in the initial year. The variance of the girls' improvements is lower. Girls at each performance level are more likely to drop out of participating. And girls are underrepresented among the high-scoring entrants. To clarify the relative importance of the various differences in the dynamics, we propose a method for decomposing the net change in the fraction female among high scoring students into several components. The decomposition suggests that the most important gender-related difference is that fewer girls are making large enough increases from year to year to move up into the top rank groups.

Section 5 ventures into the realm of assessing potential explanations for the gender gap in a more causal-inference style, looking at whether a portion of the gap may be attributable to gender-related differences in students' reactions to disappointment. Specifically, we note that the structure of the AMC contests is such that high-achieving students will be quite disappointed if they fall short of a threshold score needed to move on to a second stage exam, and that this disappointment can be viewed as a treatment that is applied at a different cutoff level of performance on different tests. We use a variant of a regression-discontinuity design to examine a narrow window around the cutoff for progressing to the second stage exam and find strong evidence that both boys and girls are more likely to drop out of participating in future years if they score just below the cutoff. We also find that the tendency to drop out after experiencing disappointment may be more common among girls.

The final Section provides a more complete recap of results and presents conclusions and implications for future research.

[6]These differences in preferences have real-world implications – Buser et al. (2014) analyze data on Dutch high schoolers and find that, although boys and girls display similar levels of academic ability, boys choose substantially more prestigious math- and science-intensive academic tracks, and that the gender difference in competitiveness accounts for a substantial portion (about 20 percent) of the gender difference in track choice. In a related finding, Azmat et al. (2016) show that gender differentials in the performance of Barcelona high school students on standardized tests depends on the stakes of the tests for university entry. Hogarth et al. (2012) study gender differences in a TV game show testing general knowledge, and show that women earn 40 percent less than men and exit the game prematurely at a faster rate.

[7]See, e.g., Chachra et al. (2009) on the extracurricular activities of engineering students.

Our investigation is related to a number of literatures. As noted above, a number of papers including Hedges and Nowell (1995), Guiso et al. (2008), and Ellison and Swanson (2010) have noted that girls are underrepresented among the high scorers on standard math assessments and math contests both in the US and in many other countries. Relative to this literature, we add a number of new observations about the gender gap. This includes both our basic observations that extreme gender gaps among the very highest achievers like those noted by Hyde et al. (2008) and Ellison and Swanson (2010) are already present by 9th grade and that the gender gap among high achievers widens over the course of high school, and our many observations about the dynamics of achievement among high-achieving boys and girls.

Our paper is also related to the literature on gender-differences in attitudes toward competition. Niederle and Vesterlund (2007) and Niederle et al. (2013) found a clear gender gap in willingness to enter contests in laboratory experiments. Prior experimental evidence has also demonstrated that men and women react differently to losing contests: Gill and Prowse (2014) find that women who lose a contest score lower in subsequent contests; Buser (2016) finds that men (but not women) react to losing by seeking greater challenges; and Buser and Yuan (2018) find that, even within populations who have already opted into competing, women are more likely to react to losing by ceasing to compete. Our real world evidence on students' reactions to disappointment are consistent with there being a similar gender gap, although our message is not entirely aligned in that we find that boys are also reacting to disappointment by dropping out of future competition.[8]

Our Section 5 analysis is very closely related to Buser and Yuan (2018) and Iriberri and ReyBiel (2018), which consider gender gaps in progression within multi-stage competitions. In particular, Section 4 of Buser and Yuan (2018) contains a regression discontinuity analysis on future participation decisions of students who are near the threshold for advancing to the second round of the Dutch Math Olympiad. They estimate that girls failing to advance are about 11 percentage points more likely to drop out of participating in the next year, and boys one percent more likely, but with standard errors of about 6 percentage points, the effect on girls is only barely statistically significant and they cannot say whether there is any effect on boys or whether the male-female difference is significant.[9] Our much larger

[8] Our finding on the magnitude of the 9th grade gender gap can also be seen as suggestive that differences in interest in competition are producing part of the real-world effect of girls being underrepresented among the highest scorers on the contests.

[9] Our work is somewhat different from Iriberri and ReyBiel (2018) in that they consider performance across stages within a single competition, whereas we consider the effects of early disappointment on performance across years.

sample (approximately 100 times as many student-years) allows us to use narrower windows and get much more precise estimates. We find estimates of 3.4-3.7 percentage points for boys (a 12-13 percent effect, relative to dropout rates among boys scoring just above the cutoff) and 4.2-5.6 percentage points for girls (a 13-17 percent effect), with standard errors of 1.2 percentage points at most. Thus, we find suggestive evidence in support of Buser and Yuan (2018)'s finding that girls are more likely to react by dropping out, but document that the effect among boys is also substantial and that the differential effect for girls relative to boys in the United States contest is not nearly as extreme as their point estimates for the Netherlands.

More broadly, our paper is motivated in several respects by the rich literature on gender gaps in wages and career development. As summarized in Blau and Kahn (2017), gender gaps in mathematics and career-oriented college majors declined substantially between the 1960s and 1980s, but there has been less progress since.[10] The literature also contains several interesting analyses of the dynamics of the gender gap in pay and workforce participation, including Bertrand et al. (2010) and Goldin et al. (2017).

2 The High-Achievement Gender Gap in AMC Scores

In this Section, we bring out some basic facts about the gender gap among AMC high scorers. In Ellison and Swanson (2010), we noted that there was a large gender gap at high achievement levels and that the gaps have a striking pattern of being much wider at achievement levels well above those that can be reliably measured with more commonly used standardized tests. Among the new observations we make here is that the high-achievement gender gap is already quite large and has the same distinctive pattern by the time students are in 9th grade, and that the gap grows wider over the course of the high school years.

2.1 Background and data

The primary subject of our analysis is a database of scores on the Mathematical Association of America's AMC 10 and AMC 12 contests from 1999 to 2007. The tests are 25-question, multiple choice tests designed to identify and distinguish among students at very high performance levels. They are administered to over 200,000 students in about 3,000 US high

[10] Focusing on STEM fields specifically, Ceci et al. (2014) present evidence on lower female propensities to major in math-intensive subjects in college and higher female propensities to major in non-math-intensive sciences. They then examine career development in STEM fields and find greater evidence of pipeline leakage in fields such as psychology, life science, and social science, rather than in math-intensive fields in which they are more underrepresented.

schools. The AMC 10 is open to students in grades 10 and below. The AMC 12 is open to students in grade 12 and below.

Several features of the AMC exams make them well suited to studying the development of high math achievement over the high school years. One is that the tests are designed to asses a broad range of (high) performance levels and are reliable even for very high-achieving students.[11] A second is that the tests are very popular among the very best math students in the US.[12] A third appealing feature is that many of the high-achieving students in our sample take the tests annually over a four year period, which lets us track the year-to-year improvement in their absolute achievement levels. The present paper is the first to exploit this feature.

The structure of the AMC contests changed twice in the period we study. In 1999, all students took a common test similar to the AMC 12. In 2000, the AMC introduced the AMC 10 and offered younger students the option of taking either test. The AMC 10 and 12 are similar – 14 of the 25 questions were common to both tests in the first year – but to be less intimidating to younger students and less affected by knowledge of above grade-level material, the AMC 10 avoids logarithms and trigonometry, and rarely has questions as difficult as the five most difficult on the AMC 12. In 2002, the AMC began offering four tests per year: the AMC 10A and 12A were offered on one date in early February, and the AMC 10B and 12B were offered two weeks later. One motivation was to accommodate students whose school was on vacation or cancelled due to snow on the A-date. But schools could offer both the A-date and B-date tests and some students choose to take a test on each date. In 2007, about 3 percent of A-date takers also took a B-date test.

The test multiplicity necessitates rescaling scores from the various year t tests to make them comparable to other tests from the same year. In the years 2000-2006, the way in which we do this is to think of year t scores as predictors of year $t+1$ AMC 12 scores. We run separate linear regressions of year $t+1$ AMC 12 scores on scores on each year t test

[11] Ellison and Swanson (2010) note that AMC scores are a stronger predictor of how students will do when retaking the math SAT than is the previous math SAT score and the tests remain a calibrated predictor of future test scores at upper tail percentiles that are an order of magnitude higher than can be measured with the SAT.

[12] While the 3,000 AMC-offering schools is a small fraction of the total number of high schools in the US, Ellison and Swanson (2016) note the AMC disproportionately offered at the types of private and high-performing public schools attended by the students with the highest SAT scores to the extent that the majority of all presidential scholar candidates attend schools that offer the AMC. Relevant to assessing participation among extreme high achievers, they note that at least 80 percent of the highest performing students on several other math contests and mathematical research contests took the AMCs. At less rarefied achievement levels, a back-of-the-envelope calculation suggests that about 20 percent of the students at participating schools with 800s on the SAT math take the AMC contests.

and consider two year t scores to be equivalent if the predicted year $t+1$ AMC 12 score is the same. This year-ahead prediction is not possible in the final year of our data, so in 2007 we instead normalize scores by comparing the performance of students who take both an A test and a B test in 2007. Appendix A provides more details on the methodology and the resulting normalizations. An AMC 10 score of x turns out to be roughly equivalent to a score of $\frac{7}{8}x$ on the AMC 12, but there are idiosyncratic differences from test to test of about 5 to 10 points on the AMC 12's 150 point scale. There is more top-coding of AMC 10 scores than AMC 12 scores, but top-coding is still at least an order of magnitude less common than on more commonly studied standardized tests.[13]

The normalization described above is not designed to put year-t and year-t' scores on a common scale. Instead, we mostly avoid the difficulties inherent in comparing scores across calendar years by focusing on students' *ranks* within the set of students who participate in a given year. In Section 3.1 we present data which suggests that transforming scores to log ranks is a very natural way to normalize student performance in that it produces a measure in which the additive improvement in performance from year to year is similar over a wide range of (high) initial performance levels. We see the ability to renormalize scores in this way as another attractive feature of the AMC environment.

Our raw data consists of separate files of student-level scores on each test in each year. In addition to the scores, the records contain a school identifier, the state in which the school is located, an anonymized student's name, and the student's gender, grade, age, and home ZIP code. We create a student-level panel data set by merging these files assuming that two scores belong to the same student if the name and school match and the age, grade, and gender are consistent, or if the name and state are the same and the city, ZIP code, age, grade, and gender are consistent.[14]

In the full pre-2007 dataset, we match 43 percent of 9th to 11th grade students to a score in the subsequent year. Note that failures to match result both from students who do not participate in the following year and the limitations of our procedure; e.g., we will miss students who report their name differently in different years, students who skip a grade, most students who move, etc. One would expect high-achieving students to be more

[13] A perfect 150 on the AMC 10 is usually equivalent to about a 130 on the AMC 12. A few hundred students per year score at least 130 on the AMC 12 versus about 15,000 who get perfect scores on the math SAT.

[14] Only unique matches are kept in the dataset for analysis. Students' demographic variables are missing for 3-6 percent of observations; we consider two values of a variable to be "consistent" if they match *or* if one or more values is missing. Grade is considered a match between a year-t observation and a year-t' observation if $grade_t - grade_{t'} = t - t'$.

likely to take the AMC in subsequent years, and our match rates are consistent with this. For example, among 9th to 11th grade students who were among the 500 highest scoring students in their cohort, the subsequent-year match rate is 80 percent.

In our analyses of the evolution of students' scores over time, we define a student's *AdjustedScore* in year t to be the rescaling of the score that they received on the first test offered by their school. Note that, at schools that offer both the A-date and B-date tests, students who only take the B-date test in year t are coded as not participating in that year. The primary reason for this decision is that we think doing otherwise would lead to miscounts of high-scoring students.[15]

2.2 Summary statistics and the gender gap in AMC participation

In this Section we present some summary statistics on AMC scores and participation rates. Gender differences in participation rates are not large, but there is some evidence of gender-related selection into the contests that should be kept in mind in interpreting some of our results.

Table 1 provides some summary statistics on participation and scores by grade and gender. For each grade-gender pair, it reports an equally weighted average across the nine years 1999-2007 of several summary statistics for that grade and gender. For example, the 18,984 9th grade girls listed as participating indicate that this is the average number of 9th grade girls who participate in each of the nine years of our dataset. The top panel contains information for female students. Female participation grows substantially from 9th to 10th grade, from an average of about 19,000 9th grade girls per year to about 28,000 10th grade girls per year. One reason for this growth may be that some teachers may hesitate to recommend the AMC tests to 9th graders, regarding the tests as too advanced and/or too likely to be a discouraging experience. Awareness of the AMCs also presumably diffuses over time. Female participation remains roughly constant from 10th to 11th grade. It then drops by about 18 percent from 11th to 12th grade.[16] One reason for this decline may be that 12th grade scores and awards come out too late to be listed on college applications.

The bottom portion of the Table reports comparable statistics for boys. Male partici-

[15] Miscounting is a concern because most schools offer only the A-date tests and some of the most serious students will take a B-date test at another area school that offers it if their school does not. Our procedure avoids double-counting these students if the alternate location they find is a school offering the test on both dates, which we think is by far the most common situation in which this occurs. We could alternately have used all of the B-date scores with some set of matching rules to filter out potential out-of-school students. Any such procedure could at most increase the sample size by two percent.

[16] We have constructed the sample to include 9th, 10th, 11th, and 12th graders from all years, so the drop in female participation noted here should not be contaminated by the time-trend in AMC participation.

pation is about 11 percent higher than female participation in 9th grade. Its growth from 9th grade to 10th grade is similar to what we saw for girls. The series then diverge a bit more, as male participation continues to grow from 10th grade to 11th grade, and has an 11th-to-12th grade decline that is less than half as large as the decline for females in percentage terms. At the end of high school the pool of 12th grade AMC takers is about 43 percent female. While the gender gap in AMC participation increases over the course of high school and we will later investigate differential dropout rates in detail, a first takeaway is that participation rates among high-achieving girls and boys are not too different for the AMC 12. Most AMC takers presumably come from the high end of the SAT population, and as noted in Ellison and Swanson (2010), the population of students with SAT scores of 600 or above is also 43 percent female.

Table 1: Summary Statistics – Participation and Scores

Grade level	Number of Students	Statistics on *AdjustedScore*			
		Mean	St.Dev	% ≥ 100	% ≥ 120
	Girls				
Grade 9	18,984	56.8	14.8	0.7	0.04
Grade 10	28,008	60.3	15.3	1.2	0.06
Grade 11	28,348	66.3	15.7	2.9	0.11
Grade 12	23,294	69.1	16.2	4.5	0.18
	Boys				
Grade 9	21,067	61.7	16.6	2.5	0.26
Grade 10	31,152	66.0	17.2	4.0	0.40
Grade 11	33,988	72.8	17.2	7.8	0.64
Grade 12	31,391	76.0	17.8	11.3	1.04

Notes: Table reports average annual AMC participation and scores by gender and grade level.

The Table also provides summary statistics on normalized AMC scores in each grade-gender cell. The overall mean adjusted score is 66 on the AMC 12's 0 to 150 scale. We will not discuss population average scores much because the AMC tests are not a good source for insights on average performance given the highly-selected populations, but it is true that the means and variances are higher in each grade in the male population. We will say a lot about year-to-year improvement for the average AMC participant, but defer these discussions until later Sections analyzing the student-level data. Our previous papers focused on counts of students achieving scores above higher thresholds, for which we think

selection is less of an issue. Scoring 100 on the AMC 12 can be thought of as roughly similar in difficulty to scoring 780 or 800 on the math SAT. Among 12th graders scoring at this level or higher, we find a male-female ratio of about 3.4:1. The male-female ratio among students achieving comparable scores on the SATs is about 2:1. The gender gap could be somewhat different on the AMC and SAT due to differences in what is being tested and the fact that the SAT is a cruder instrument. The magnitude of the difference does suggest that there are some gender-related differences in participation rates. For example, female participation rates being a bit lower among the very highest scorers on the SAT would be expected given the literature on gender differences in attitudes toward competition.

Scoring 120 on the AMC 12 represents a much higher level of achievement – roughly in the 99.99th percentile in the full US 12th grade population. Here, we think that selection into test-taking is less important. Our primary reason for saying this is that reaching the highest levels of performance on the AMC 12 requires a great deal of both natural ability and effort directed toward mastering high school mathematics, and we feel that it is unlikely that students not interested in participating in math competitions would exert the effort necessary to excel at these levels. We see this as analogous to saying that there are unlikely to be many high school students who can throw a curveball and a 90mph fastball who are not participating in competitive baseball. We will also note later that year-to-year match rates are quite high among the highest scoring students on the AMC.[17] Note that the male-to-female ratio is much larger among students reaching this higher score level on the AMC. This is part of a larger pattern noted in Ellison and Swanson (2010). One implication is that one needs to be careful in constructing comparisons; e.g., we would expect the male-to-female ratio among 9th graders scoring at least 100 to be larger than the male-to-female ratio among 12th graders scoring at least 100, because looking at students scoring 100 in 9th grade is looking at students who are much farther out into the right tail.

2.3 The gender gap in high math achievement over the high school years

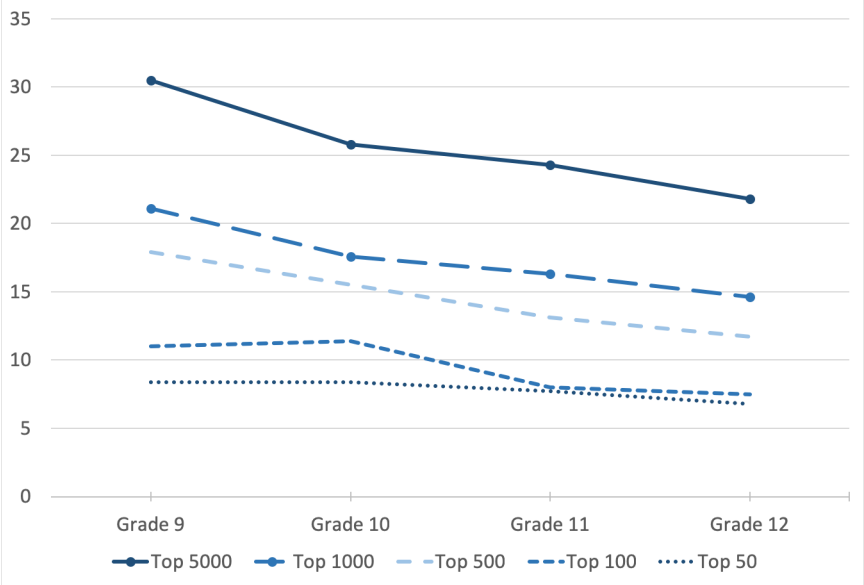
In this Section, we illustrate how the gender gap among AMC high scorers changes over the course of high school. In Ellison and Swanson (2010), we noted that there was a large gender gap among AMC high scorers and that it was larger at higher achievement levels. Among the additional observations here are that the gender gap is already quite large in 9th grade and that the gender gap widens substantially over the course of high school.

Figure 1 reports the percentage of AMC high scorers in each grade who are female for

[17] Ellison and Swanson (2016) provide some supporting evidence from examining AMC participation rates among winners of other math contests and mathematical research competitions.

various definitions of high scoring. The top line in the Figure uses the least restrictive definition of high scoring, examining the 5,000 highest-scoring students in each grade-year. These are very high-achieving students, but not extremely unusual ones: one could think of them as students on a trajectory to score 780 or 800 on the math SAT by the end of high school. At the left endpoint we see that there is a substantial gender gap in 9th grade: only 30.5 percent of the high-scoring 9th graders are female. Looking from left-to-right along this line, we see that the gender gap widens in each subsequent year. By 12th grade, only 21.8 percent of the top 5000 high-scorers are female. The drop from 9th grade to 10th grade is the largest, but the decline is fairly steady over the high school years.

Figure 1: Percent Female by Grade and Achievement Level



Notes: Figure reports the average percent female, for each achievement group, across the six cohorts that we observe for all four of their high school years.

The lower series in the Figure present comparable estimates using more and more stringent definitions of high-achieving, going all the way out to a definition that is two orders of magnitude more demanding and examines just the top 1 percent of our initial very high-achieving pool. That each of the curves slopes downward indicates that the finding that the gender gap widens over the course of high school is quite robust to how one defines high-scoring. In proportional terms, the decline in the percent female from 9th grade to 12th is between 29 percent and 35 percent along every curve except the lowest one (where

it is 19 percent).[18]

Ellison and Swanson (2010) highlighted that the gender gap is much larger when one examines more extreme high achievers; as discussed above, we argue that the largest gaps seem unlikely to be primarily due to selection into test taking. A comparison of the leftmost points of the series in Figure 1 shows clearly that this pattern is already present by 9th grade. Girls comprise 30.5 percent of the top 5000 9th graders, but only 8.4 percent of the top 50 9th graders. The decline from the top line to the bottom line is roughly similar in each grade. One implication is that it is important to understand the persistence of performance over the high school years. If performance is highly persistent (which we will find), then the Ellison and Swanson (2010) finding about how the gender gap differs for extreme high achievers relative to ordinary high achievers cannot be primarily a finding about things that are happening during high school. However, the gender gap widens over the high school years among ordinary high achievers, extreme high achievers, and everyone in between. Our subsequent analyses are motivated in large part by a desire to better understand this widening of the gender gap.

3 Dynamics of Achievement Among High Achievers

In this Section, we take a step back from gender-related issues and present some more general evidence on the dynamics of achievement among high-achieving math students. Among our observations are that the distribution of mathematical achievement is sufficiently spread out so that the top 9th graders are already very high up in the overall score distribution, that high-achieving students must substantially improve their performance from year to year to keep up with their cohort, and that there is substantial performance persistence: the highest scoring students are much more likely to achieve a very high score in the following year than students ranked just a little lower, and it is unlikely that students will greatly improve their within-cohort rank.

3.1 Growth and variation in absolute performance

Although it is becoming increasingly common to take calculus in the junior year and the AMC contests only cover precalculus topics, top students are increasing their command of the AMC material and problem-solving techniques over the course of high school.[19] To give

[18]The latter estimate is fairly noisy given the small sample sizes: the top 50 is only 7-8 percent female, which means that there are typically just 3 or 4 girls in the top 50 of each grade-year.

[19]In 2015, over 120,000 AP Calculus exams were taken by students in 11th grade and below. It was less common for the cohorts we study, but there were already over 30,000 students in 11th grade or below taking

some sense of how performance grows over time, Table 2 lists the average overall rank that a student needed to have in order to be in the grade-specific top 50, top 100, top 500, etc. For example, to rank among the top 100 9th graders, one only needs to score in the top 1,173 overall, whereas a 12th grader needs to score in the top 241 overall to be in the top 100 in his or her cohort.

Table 2: Growth in Absolute Performance

Within-grade rank	Corresponding overall rank				Decrease in overall rank to maintain rank in grade		
	Grade 9	Grade 10	Grade 11	Grade 12	$9 \rightarrow 10$	$10 \rightarrow 11$	$11 \rightarrow 12$
5000	52,554	32,686	15,654	11,395	38%	52%	27%
1000	15,674	5,734	3,293	2,186	63%	43%	34%
500	8,350	3,234	1,738	1,356	61%	46%	22%
100	1,173	668	290	241	43%	57%	17%
50	875	310	152	106	65%	51%	30%

Notes: Table reports the full-population rank of the Nth-best student in each grade.

One immediate observation from the Table is that some students have already reached very high achievement levels by 9th grade. For example, the 500th best 9th grader is already well within the top 5000 12th graders, and hence is already at the level where we would expect a nearly perfect SAT score. The 50th best 9th grader is similarly well within the top 500 12th graders.

While some 9th graders are already very good, the Table also makes very clear that students must improve substantially from year to year if they wish to maintain their within-cohort position. The right panel reports the percentage reduction in the overall rank that students in various positions must make to maintain their within-grade rank. High scoring 9th graders will need to improve their overall rank by roughly 40-60 percent in order to achieve the same position relative to their peers as a 10th grader. High scoring 10th graders will need to improve their overall rank by about 50 percent. The required improvement between 11th and 12th grades is somewhat smaller. But we think of the 20-30 percent improvement required as still surprisingly large, given that the 12th grade competitor pool is smaller and most high scoring 12th graders will be studying calculus (or something more advanced), which is not covered on the AMC tests.

The similarity of the percentage change numbers within each column is striking given

AP Calculus when our first cohort was in 11th grade (2001).

that the stringency of the definition of high achievement varies by two orders of magnitude from the top to the bottom. We take this as suggesting that the log of a student's rank is a natural cardinal measure of performance to use when analyzing high-achieving students. We see this as another feature of the AMC environment which makes it attractive to study.[20]

One simple way to get a feel for what size of year-to-year improvement is typical at the individual level is to examine the distribution of $\log(Rank_{t+1}) - \log(Rank_t)$ among students who take the test in both years t and $t+1$. This variable has a mean of -0.28 for 9th graders, -0.39 for 10th graders, and -0.26 for 11th graders. These are substantial increases in performance. For example, a student who scores at the 50th percentile of the AMC-taking population as a 9th grader and improves his or her log rank by the expected amount in each year would reach the 62nd percentile in 10th grade, the 75th percentile in 11th grade, and the 80th percentile in 12th grade. However, note also that they are not nearly enough to bring the already strong 9th grader in the example up to the top of the distribution by the end of 12th grade. We think of this as another way in which AMC scores suggest that the distribution of mathematical performance is quite spread out even among students that are normally lumped together as high achievers.

One would expect that the degree to which students improve from year to year will differ for students in different parts of the distribution. The effort that students are putting into improving their knowledge and problem solving skills will differ. And although the right panel of Table 2 suggests that a log-rank transformation of performance is natural, any cardinalization of mathematical achievement is inherently arbitrary. Because AMC performance in any given year is a noisy measure of a student's underlying achievement level, one cannot estimate average achievement gains as a function of initial achievement via an OLS regression. We can, however, use IV regressions to estimate such gains when some instrument for the measurement error is available. The columns of Table 3 present regressions of $\log(Rank_{t+1}) - \log(Rank_t)$ on $-(\log(WithinGradeRank_t) - \log(5000))$, using the log of a student's within-grade rank in year-$t-1$ as an instrument on the subsample where this variable is available. The constant terms in these regressions can be thought of as the average improvement for a student who has the 5000th best score in their cohort in year t. The estimates suggest that these improvements in log rank are -0.50 for 10th graders and -0.33 for 11th graders. If we convert the mean improvements in $Rank$ needed

[20] When student performance can only be measured as a within-year z-score, the dynamics of the year-to-year changes in relative-to-cohort performance are more difficult to analyze for high-scoring students because changes are highly asymmetric: high-achieving students can only improve their performance very slightly from year to year, but can easily do much worse.

to maintain a given within-grade rank in Table 2 to changes in $\log(Rank)$, they would be approximately -0.74 in 10th grade and -0.32 in 11th grade. Hence, a 10th grader who ranks 5000th in his grade must improve by substantially more than the expected amount in order to maintain his or her rank. Intuitively, this reflects that there are many more students ranked below the 5000th student than above. If the 5000th ranked student makes the average improvement, then there will be more students jumping ahead of her due to above-average gains than falling behind due to below-average gains.

Table 3: Average Achievement Gains as a Function of Initial Achievement

	Dep. Var.: $\log(Rank_{t+1}) - \log(Rank_t)$			
	$10th \rightarrow 11th$		$11th \rightarrow 12th$	
Variable	Coef.	Std. Err.	Coef.	Std. Err.
Constant	-0.50***	(0.005)	-0.33***	(0.005)
$-(\log(GradeRank_t) - \log(5000))$	-0.07***	(0.004)	-0.05***	(0.004)
Number of observations	81,430		100,270	
Root MSE	0.92		1.01	

Notes: Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Table reports the results of IV regressions of growth in absolute performance as a function of initial performance relative to cohort. $\log(GradeRank_t)$ instrumented with $\log(GradeRank_{t-1})$.

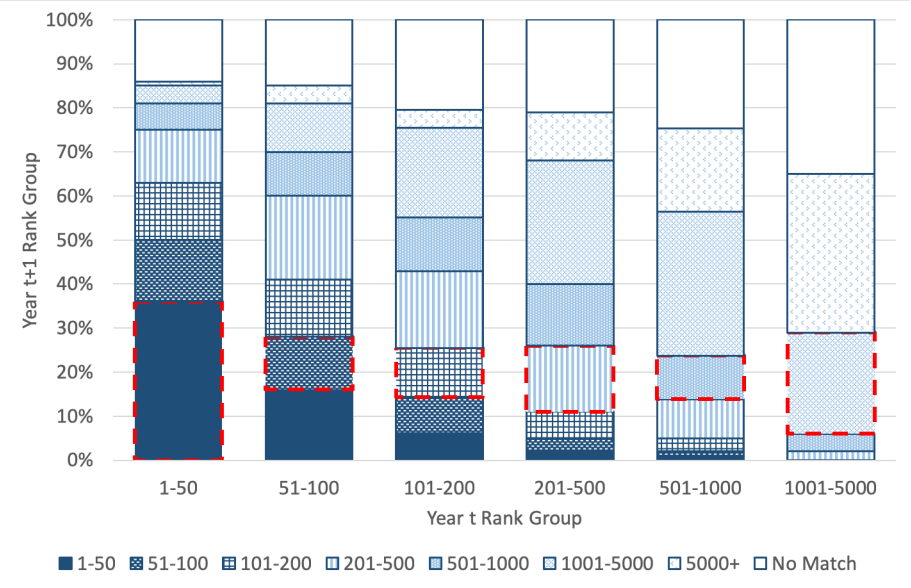
The negative coefficient estimates on the term reflecting initial achievement levels indicates that students at higher achievement levels in the initial year are expected to make even larger improvements in log rank. For example, the predicted improvements for a student who ranks 500th in his or her cohort in year t are -0.66 for a 10th grader and -0.45 for an 11th grader. The former is very close to what is needed for a 10th grader to maintain the same rank; the latter is more than sufficient for an 11th grader.

Standard deviations of the full sample increases in log rank are 0.73, 0.86, and 0.96 for 9th to 10th, 10th to 11th, and 11th to 12th grades, respectively. Note that these will reflect both the measurement error of the test as a measure of students' underlying achievement levels in both years, and also true variation in the growth in achievement from year to year. Appendix B presents a calculation examining changes over multiple years to estimate the relative importance of the two components. It suggests that the measurement error component is larger than the variation in achievement growth component, but that there is still substantial heterogeneity in students' true achievement from year to year.

3.2 Persistence and mobility in relative-to-cohort performance

We now focus on how students move up and down *within their cohort* from year to year. Figure 2 presents a graphical view of the estimated rank-to-rank transition matrix. For example, the height of the darkest shaded portion at the bottom of the left-most bar indicates that there is a 36 percent chance that a student who is among the top 50 in their cohort in year t will again rank in the top 50 in year $t+1$, and the portion of the same bar just above this indicates that there is an additional 16 percent chance that such a student will rank from 51 to 100 in year $t+1$.[21]

Figure 2: Persistence in Math Performance – Forward Transition Matrix



Notes: Figure reports the forward transition probability of each year-$t+1$ within-grade rank group for students in each year-t within-grade rank group.

One clear observation from the Figure is that performance in year t is a strikingly strong predictor of performance in year $t+1$, even when making comparisons that rely on fine distinctions in year-t performance. Comparing students who were ranked in the top 50 in

[21] Due to the discreteness of AMC scores, there will typically be a number of students tied for positions that cross each boundary. For example, in 2006, fourteen 11th graders had scores of 124, which left them tied for positions 196 to 209. In this situation, we would include the experience of each of these students with weight 0.64 in our calculation of what happened to students with ranks of 201 to 500 in year t. And we similarly record each student's outcome as their probability of being in each rank group as though ties are broken at random.

their grade in year t to those ranked 51-100, for example, the higher-ranked students are more than twice as likely to achieve a top 50 score in year $t+1$ (36 percent vs. 16 percent), and less than half as likely to score outside the top 500 (10 percent vs. 25 percent). Similar patterns are visible over and over in the other bars. Students who were ranked from 51-100 are more than twice as likely to achieve a top 100 score in year $t+1$ than are students who were ranked 101-200 at t. Students ranked from 101-200 at t are more than twice as likely to achieve a top 200 score at $t+1$ than are students who ranked 201-500, and so on.

A second observation is that it is possible to move up in the distribution, but declines are much more common and substantial improvements are quite unlikely. To help visualize this, we have outlined boxes that correspond to the diagonal of the transition matrix using red dashed lines. The fact that much of each bar is above the diagonal box is an illustration of the frequency of declines in within-cohort ranks. For example, 53 percent of students ranked from 201-500 in year t will score outside the top 500 in year $t+1$. Some substantial improvements are also visible in the Figure. For example, 14 percent of those ranked 101-200 within their grade in year t move into the top 100 in year $t+1$, including some moving into the top 50. But the chance of moving up by even one rank group is never above 16 percent and the chances of all of the three-or-more group improvements are sufficiently small as to be very hard to make out in the Figure.

A third observation is that dropping out of participation is relevant even among high-achieving students. The heights of the white outlined boxes at the top of each bar correspond to the percentage of students whom we were not able to find in the year-$t+1$ data. In our full sample, we are unable to match 57 percent of grade 9-11 year-t participants to a year-$t+1$ score. Among students who are ranked from 1001-5000 in their grade in year t, the percent unmatched drops to 35 percent. But the fact that the unmatched rate is 35 percent for students with ranks from 1001-5000 and just 14 percent for students with ranks from 1-50 suggests that at least 20 percent of the students in the 1001-5000 truly do not participate in year t. Dropping out appears to be less and less likely as one moves up in the ranks. The majority of the unmatched students in the top group are probably unmatched because of the limitations of our dataset rather than due to the students actually dropping out.[22]

[22]To investigate this issue, we looked manually through published lists of 2006 and 2007 high scorers. Among the top 50 students in each grade in 2006, we failed to find 2007 matches for 2.6 percent of 9th graders, 4.3 percent of 10th graders, and 12.1 percent of 11th graders. These figures should be compared to the sum of the dropout rate and the probability of finishing outside the 2000 in our algorithmic match, which is about 18 percent on average across grades. Several factors are involved in the superiority of this manual match over our algorithmic match: manually, we were able to identify students who switched schools, students who took the test at a testing center in one year and in their high school in another year, and

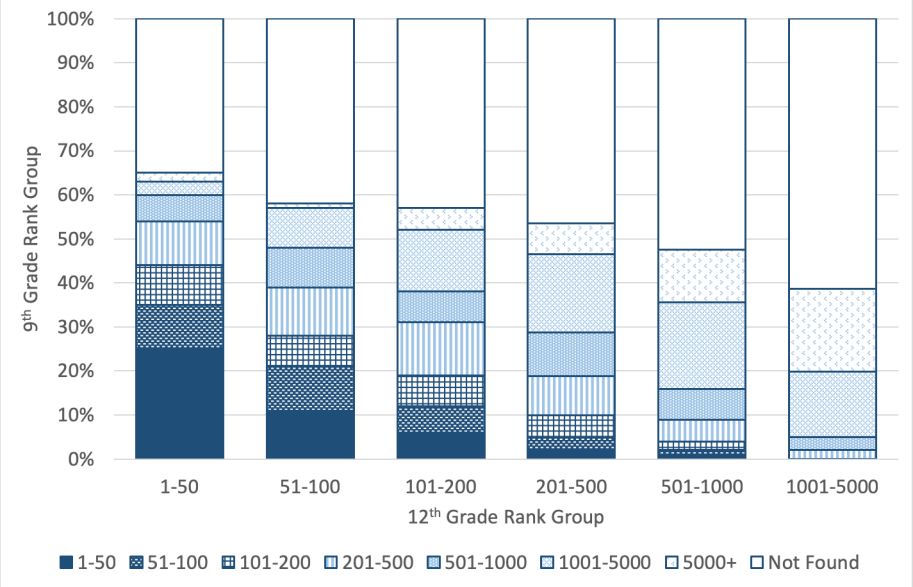
One final comment on the Figure is that we feel it bolsters the case that the AMC is an interesting measurement tool. While we always encourage readers to look up old test questions online, with the belief that many will feel that the test seems nicely designed to test both problem solving skills and students' command of core precalculus topics, such impressions cannot tell us how noisy a test is as a measure of some student capability, nor how much we should care about the capability being measured. In the case of the AMC, the level of persistence in Figure 2 makes very clear that the test is a sufficiently accurate and consistent measure of some capability related to high achievement such that it is a good predictor of year-ahead performance. And our earlier results on students' gains from year to year indicate that the capability being measured is something that builds over the high school years, versus something more stable like differential quickness or accuracy in performing calculations.

We also present here a longer horizon backward-looking transition matrix. The bars in Figure 3 show the fraction of students who achieved the rank corresponding to that bar in 12th grade who were in each rank category in 9th grade. At the very highest levels of achievement, the performance persistence we noted earlier remains striking. For example, we can see in the first bar that there are more holdovers from the 9th-grade top 50 in the 12th-grade top 50 (about 25 percent) than there are students who have moved up from the entire 201-40,000 range (about 21 percent). Only 5 percent scored outside the top 1000 as 9th graders. Although there are a substantial fraction, 35 percent, whom we were unable to match to a 9th grade score, given how few students manage to move up from the 1000+ range into the top 50, we imagine that many of these students are students whom we failed to match rather than true entrants. Some causes of matching failures, including students who switch high schools or skip grades, will likely be more frequent here as we are matching across a three year span.

At the still extremely high level of students who rank 201-500 among 12th graders, there is more heterogeneity in 9th grade origins. Students moving down from the top 200, holdovers from the 9th grade 201-500 group, and students moving up from the 501-1000 group each comprise about 10 percent of this group. We also see a much larger number of

students who appear to have listed their first name differently in different years. While the 12 percent dropout rate after 11th grade may seem surprising, it includes at least one extremely strong student who left high school after 11th grade to start college, as well as several 11th-to-12th grade dropouts whose 2006 AMC scores were surprisingly high given their previous 2005 AMC scores and their subsequent 2006 AIME scores. It is worth noting that matching failures are likely more prevalent at the highest score levels due to high-performing students taking the exams at testing centers in lieu of or in addition to their own high schools.

Figure 3: Early Performance of Top Math Students – Backward Transition Matrix



Notes: Figure reports the probability of each within-9th-grade rank group for students in each within-12th-grade rank group.

students who had not done as well in 9th grade, with 25 percent coming from outside the top 1000 ranges.

At the lower (but still high) levels of 12th grade achievement in the Figure, improvement since 9th grade plays an even more prominent role. Only about 5-9 percent of these students in the 12th-grade 501-1000 and 1001-5000 rank groups are students who have dropped down from a higher 9th grade rank group. Meanwhile, 12 percent and 19 percent, respectively, are students who have moved into these groups after having scores that placed them outside the top 5000 9th graders. These students have improved by enough to overcome both their initial disadvantage and the substantially higher score needed to make the within-grade top 5000 as a 12th grader. The fraction of students that we cannot match to a 9th grade score is also much larger in these groups at 53 percent and 62 percent, respectively. The fact that the failure-to-match rate is so much larger here than it was for the top 50 students suggests that a substantial number of the unmatched 12th graders in these groups are true entrants who had not participated in 9th grade.

Early in this Section, we noted that the gender gap among high-achieving math students

is already large in 9th grade. Given that performance is highly persistent, it is not surprising that the girls are not able to overcome their initial disadvantage. But performance persistence makes it all the more striking that the gender gap among high-achieving math students actually widens substantially over the the high school years. Some of the more detailed findings in this Section highlight channels that could be relevant: large performance improvements are needed to maintain one's within-cohort rank; some students are dropping out of participating (at least at all but the highest ranks); and the three-year time span between 9th and 12th grades is long enough to allow quite a number of students who were not high-performers in 9th grade to improve or enter and achieve a high rank by the end of high school. Gender-related differences in any of these dimensions could contribute to the widening gender gap that we document.

4 Gender Differences in Dynamics and a Decomposition

In this Section, we look at gender-related differences in the dynamics of year-to-year performance and present a decomposition that lets us quantify the relative importance of several factors to the broadening of the gender gap in high math achievement over the high school years.

4.1 Differences in dynamics

A number of gender-related differences could in principle lead to a widening gap: girls might be more likely to drop out of participating; participating girls might improve less on average from year to year; there may be less variance in year-to-year improvement for girls; and/or fewer high-achieving girls may drop into participating.

We first look for gender-related differences in year-to-year improvement within the population of students who participate in the AMC tests in consecutive years. Table 4 presents estimates from an OLS regression,

$$\begin{aligned}\log(\text{GradeRank}_{it+1}) - \log(\text{GradeRank}_{it}) \;=\; & \beta_1 \text{Female}_i + \beta_2 \log(\text{GradeRank}_{it}) + \\ & \beta_3 \log(\text{GradeRank}_{it})^2 + \beta_4 \text{Female}_i \times \log(\text{GradeRank}_{it}) + \\ & \beta_5 \text{Female}_i \times \log(\text{GradeRank}_{it})^2 + \\ & \beta_6 \text{B-Date}_{it} + \beta_7 \text{Both}_{it} + \delta_{g(it)} + \gamma_t + \epsilon_{it}\end{aligned}$$

where the δ_g and γ_t are grade and year dummies. Note that the dependent variable is the increase in a student's rank, so that a positive coefficient corresponds to a smaller increase

in performance. The left panel reports estimates from a regression run on the set of students who ranked in the top 5000 within their grade in the initial year.[23] The negative coefficient on the initial rank indicates substantial mean-reversion in within-grade rank, as one would expect given that test scores are a noisy measure of underlying ability.

The primary coefficient of interest in the regression is the coefficient on the Female dummy. It is positive and highly significant, indicating that girls are improving by less from year to year than boys by about 31 log points. The second main estimate of interest is whether there are gender-related differences in the variance of year-to-year improvement. The lower panel of the Table reports gender-specific means of the squared residuals from the above regression. Again, we find a statistically significant gender difference: there is greater year-to-year variance in the boys' performances. Hence, we have identified two separate features of the dynamics that would tend to contribute to a widening of the gender gap among the highest achievers: (1) the girls' mean improvement from year to year is lower; and (2) the variance in their year-to-year improvement is also lower.

In the above regression, there is also a moderately-sized but statistically significant coefficient on the interaction between the Female dummy and within-grade rank, indicating that the gender gap in mean improvement is larger for higher achievers. To examine whether this may reflect a substantial difference among the highest achievers, the right panel of Table 4 estimates the same regression on the sample of even higher achievers who were ranked in the top 500 in their cohort in the initial year. We find that things are not appreciably different at this level. The gender gap in mean improvement is estimated to be 32 log points per year, and the residual variance is again lower for the girls. In unreported results, we also estimated the above regressions separately on 9th, 10th, and 11th graders and did not find substantial differences in either finding across grades.

Recall that we earlier noted that the fraction of year-t students whom we cannot match to a year-$t+1$ score is substantially higher for students lower down in the top 5000 than for the highest scorers, which suggested that dropping out of test-taking is relevant to the composition of the set of high scorers. To look for gender-related differences in dropout rates, we define an indicator Dropout_{it+1} for whether each year-t high scorer could not be

[23]The sample is restricted to the set of students who were in grades 9, 10, or 11 in the initial year and whose genders were non-missing. The regressions also include unreported year and grade dummies, dummies for students taking the B-test, and dummies for students taking both the A-test and B-test. The latter variables are intended to control for unobserved differences in students' commitment to the contests. Lastly, the $\log(\text{GradeRank}_t)$ control is adjusted by subtracting the sample mean.

Table 4: Gender Differences in Within-Cohort Rank Dynamics for High-Achieving Students

	Dep. Var.: $\log(\text{GradeRank}_{t+1}) - \log(\text{GradeRank}_t)$	
Variable	Top 5000 in grade at t	Top 500 in grade at t
Female	0.31***	0.32***
	(0.012)	(0.055)
Adj $\log(\text{GradeRank}_t)$	-0.25***	-0.09***
	(0.007)	(0.024)
Adj $\log(\text{GradeRank}_t)^2$	-0.02***	0.01
	(0.002)	(0.010)
Female $\times$ Adj $\log(\text{GradeRank}_t)$	-0.03*	0.01
	(0.014)	(0.065)
Female $\times$ Adj $\log(\text{GradeRank}_t)^2$	-0.01	0.01
	(0.007)	(0.039)
$\hat{\sigma}^2_{\text{male}}$	1.51***	2.18***
	(0.009)	(0.033)
$\hat{\sigma}^2_{\text{female}}$	1.20***	1.99***
	(0.015)	(0.078)
Number of observations	81,570	9,682

Notes: Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

found in the year-$t+1$ data, and estimate the OLS regression

$$\begin{aligned}\text{Dropout}_{it+1} &= \beta_1 \text{Female}_i + \beta_2 \log(\text{GradeRank}_{it}) + \beta_3 \log(\text{GradeRank}_{it})^2 + \\ &\quad \beta_4 \text{Female}_i \times \log(\text{GradeRank}_{it}) + \beta_5 \text{Female}_i \times \log(\text{GradeRank}_{it})^2 + \\ &\quad \beta_6 \text{B-Date}_{it} + \beta_7 \text{Both}_{it} + \delta_{g(it)} + \gamma_t + \epsilon_{it}\end{aligned}$$

Recall that Dropout_{it+1} will reflect both true dropouts and students whom we fail to match due to inconsistently reported names, etc.[24]

The first column of Table 5 reports estimates from a linear regression run on the full sample of 9th-11th grade students who were in the top 5000 in their grade in year t. The mean dropout rate in this sample is 32 percent, and dropout rates are similar across grades. The primary coefficient of interest is the Female dummy. The estimate of 0.023 indicates that girls are 2.3 percentage points more likely to drop out of participating than boys

[24]The B-test dummy takes on a value of 0.02 and is statistically significant. We suspect that this reflects in part that a higher fraction of students taking B-date tests are students taking the test at a location other than their regular school, which makes us more likely to fail to match their performances across years. We hope that such matching failures are not gender-related.

with comparable scores. Again, the estimate is highly statistically significant, so we have identified a third factor contributing to the widening of the gender gap over the course of high school.

Table 5: Gender Differences in Dropout Rates for High-Achieving Students

	Dep. Var.: Dropout $t \to t+1$				
	Sample: Top 5000 in grade X in year t				Top 500
Variable	All grades	Grade 9	Grade 10	Grade 11	All Grades
Female	0.023***	0.005	0.021***	0.045***	0.002
	(0.004)	(0.006)	(0.006)	(0.007)	(0.013)
Adj log(GradeRank$_t$)	0.069***	0.073***	0.066***	0.069***	0.030***
	(0.002)	(0.004)	(0.004)	(0.004)	(0.006)
Adj log(GradeRank$_t$)^2	0.008***	0.008***	0.009***	0.008***	0.004
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Female × Adj log(GradeRank$_t$)	0.001	0.006	-0.004	-0.002	-0.005
	(0.005)	(0.008)	(0.008)	(0.008)	(0.016)
Female × Adj log(GradeRank$_t$)^2	0.001	0.004	-0.005	0.001	0.007
	(0.002)	(0.004)	(0.004)	(0.004)	(0.008)
Number of observations	119,325	39,284	39,747	40,294	12,020

Notes: Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

The second through fourth columns of the Table present similar regressions estimated separately on the students in 9th, 10th, and 11th grades, respectively. Here, we do see a substantial difference across grades. The gender gap in dropout rates is much larger in the 11th to 12th grade transition than in the other years. We estimate that girls are 4.5 percentage points less likely to participate in 12th grade than boys who had comparable 11th grade scores. Early in high school, the gender gap in dropout rates is much smaller.

All regressions include controls for the student's within-grade rank in the initial year.[25] The coefficients on these controls reflect that higher-scoring students are substantially less likely to drop out. The coefficients are quite similar across all three grades, indicating that this relationship is fairly stable over the course of high school.

The final column of Table 5 looks at more extreme high scorers who were among the top 500 students in their grade in year t. The mean dropout rate in this sample is lower at 19 percent, reflecting again that higher-scoring students are less likely to drop out in year

[25] We have normalized this variable separately by grade to have mean zero within the sample of students in each grade, to facilitate interpretation of the coefficient of the Female dummy as reflecting the difference in dropout rates for the mean student in our sample.

$t+1$. The point estimate of the gender-related difference in dropout rates is also much smaller, just 0.2 percentage points, but the standard error is such that we can neither reject that the gender gap is zero, nor that it is the same as in the top 5000 sample – as noted in Figure 1, only 12-18 percent of these extreme high scorers are female, so this coefficient is estimated on a relatively small sample. In unreported grade by grade regressions, the gender gap in dropout rates again appears much larger in 11th grade than in the earlier years, but the standard errors are such that the only statistically significant conclusion one could draw about top 500 students is that that there is a substantial gender gap in dropout rates after the 11th grade year.

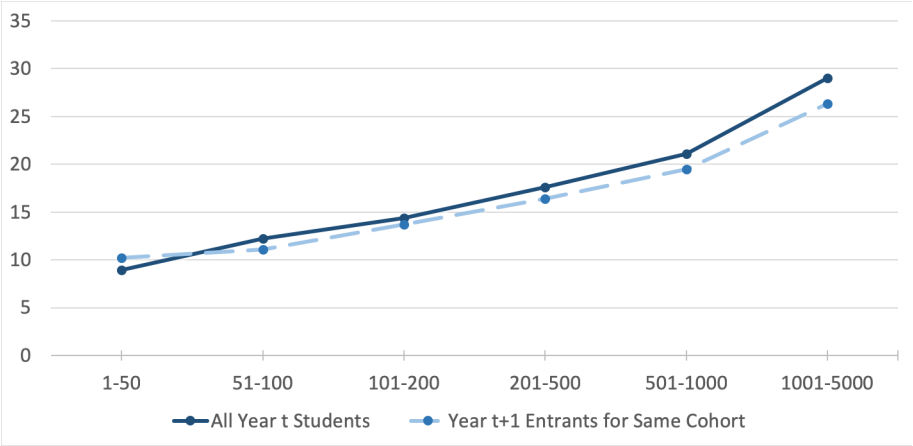
We noted earlier that a nontrivial number of high scorers at the end of high school are students whom we could not match to a 9th grade score. While some of this is due to difficulties in matching, part is the real phenomenon of students participating in the AMCs after not having been involved with math competitions from the outset. To examine whether there are also gender-related differences in this aspect of the dynamics, Figure 4 graphs the fraction female among all grade 9-11 students who were in each rank group in some year from 1999-2006, and the fraction female among grade 10-12 students in the rank group in 2000-2007 who are entrants; i.e., students who are in the rank group in year $t+1$ and whom we were unable to match to the year-t dataset.[26] In all but the top rank group, we find that the fraction of female students among the entrants is slightly lower than the fraction among the students who were in that group in the previous year. On average, the difference is about one percentage point.[27] This gender difference in AMC entry is a fourth contributor to the broadening of the gender gap over the high school years.

To recap, we have identified several gender-related differences in the dynamics of student achievement that will contribute to the widening of the gender gap in high achievement on the AMC over the high school years. High-achieving girls are on average not improving by as much from year to year, there is less variance in their year-to-year improvement, they are more likely to drop out of participating (especially after 11th grade), and we see fewer girls among the high-scoring entrants whom we cannot find in the previous year's data.

[26] All figures are simple unweighted averages of the means for each grade-year cell.

[27] It is possible that there are gender-related differences in our ability to match students; e.g., one gender could be more likely to fill in their name differently in different years. However, girls are overrepresented in the pool of year-t high scorers whom we cannot match to a year-$t+1$ score, and underrepresented among year-$t+1$ high scorers whom we cannot match to a year-t score: the potential gender-related matching errors suggested by these results have opposite sign.

Figure 4: Gender Composition of AMC Entry



Notes: Figure reports the gender composition of highly ranked students new to the AMC in comparison with student in the rank group in the previous year.

4.2 A decomposition of changes in the gender gap

In the previous Section, we noted that the dynamics of boys' and girls' achievement differ in multiple ways. In this Section, we define a decomposition of the change in the gender gap into portions attributable to various differences that provides a measure of their relative importance.

Our analysis focuses on changes in the fraction μ^f_{Xt} of students in achievement group X at time t who are female. (We will often use being in the top 50, 500, or 5000 as the group X.) Here, we relate this to various aspects of differences in the boys' and girls' transition matrices.

Proposition 1 *The change in the fraction female in group X can be written as:*

$$\mu^f_{Xt+1} - \mu^f_{Xt} = \Delta^{drop}_X + \Delta^{cont}_X + \Delta^{grow}_X + \Delta^{entry}_X + \Delta^{mech}_X.$$

See Appendix C for algebraic expressions of each term and proof.

The first term in the decomposition, Δ^{drop}_X, can be thought of as the change in female representation that is due to girls dropping out at a different rate (assuming that the girls who dropped out would have succeeded at the same rate as the girls who continued to participate). The second term, Δ^{cont}_X, reflects the difference in rates at which girls who continue to participate improve by enough to remain in rank group X. The third term, Δ^{grow}_X,

reflects the difference in rates at which lower-ranked girls versus boys subsequently climb into group X. The fourth, Δ_X^{entry}, reflects any discrepancies between female representation among the high scorers who did not participate in the previous year and what would be expected given the total number of entrants and female representation among the previous year's high scorers.

The final term in the decomposition, Δ_X^{mech}, captures mechanical changes that would occur even if there were no gender-related differences in the transition process, due to asymmetries in the initial conditions. There are mechanical effects pushing in both directions. A negative effect is that the girls in each rank group X are disproportionately found in the lower part of the rank group, so girls in X would be less likely to avoid dropping into a lower group in the following year. Working in the opposite direction, there are also more girls in the rank group just below X than in group X. With gender-independent dynamics, this would result in the set of students who move up into rank group X in the next year being more heavily female.The sign of the net mechanical effect Δ_X^{mech} will depend on which of these countervailing effects is larger.

As discussed in further detail in Appendix C, we implement this decomposition by estimating the transition probabilities both for the full population and for girls as smooth functions of the initial year rank via local linear regressions, with $\log(Rank)$ as the right-hand-side variable. We do this separately for students in 9th, 10th, and 11th grades, pooling the data for all six cohorts within each regression.

One version of our basic fact about the widening gender gap was that the percentage of female students in the top 5000 drops from 30.5 in 9th grade to 21.8 in 12th grade. This is a drop of 8.7 percentage points over three years, which is about 3 percentage points per year. The first row of Table 6 presents a decomposition of this change.[28] It indicates that by far the largest source of the drop – indeed responsible for 3.6 percentage points, which is more than 100 percent of the drop – is Δ_X^{grow}, the term in our decomposition which reflects differences in the rates at which male and female students at ranks below 5000 improve their performance and "grow" into the top 5000. Note that this term is designed to control for how far below the top 5000 cutoff male and female students were in the previous year: it is due only to differences in the probabilities that male and female students at each given rank outside the top 5000 move up into the top 5000. This in turn will reflect both the differences we identified earlier in both average improvements from year to year and in the

[28]The "average" decomposition is obtained by averaging separately estimated decompositions of the changes from 9th to 10th, 10th to 11th, and 11th to 12th grades.

variance of students' improvements.[29]

Two other features of the dynamics are a little less than one-third as important as the growth effect: Δ_X^{cont} which reflects the reduced rate at which highly-ranked female students who take the test maintain their top 5000 position; and Δ_X^{entry} which reflects the lower fraction of female students among "entrant" high scorers. The difference in dropout rates is a smaller contributor on average.

Table 6: Decomposition of Declines in Fraction Female in Top Rank Groups

Grade Level	Achievement Level	Change in % Female	Decomposition of decline				
			Drop	Cont	Grow	Entry	Mech
Average	Top 5000	-3.1	-0.4	-1.2	-3.6	-1.1	3.5
$9 \rightarrow 10$	Top 5000	-4.6	-0.1	-1.4	-2.8	-2.2	1.5
$10 \rightarrow 11$	Top 5000	-2.3	-0.4	-1.1	-3.7	-0.8	3.6
$11 \rightarrow 12$	Top 5000	-2.1	-0.9	-1.1	-4.3	-0.3	4.7
Average	Top 500	-2.0	-0.3	-0.9	-4.5	-0.4	3.8
Average	Top 50	-0.5	-0.3	-0.8	-3.0	0.2	3.2

The final column indicates that the total drop would be much larger were it not for a positive mechanical effect: the growth-related subcomponent turns out to be much more important than the continuation-related subcomponent. To appreciate why this effect can be large in practice, recall that the fraction female is much higher in the population of test-takers outside the top 5000. For example, for 10th graders it is 0.26 for students in the top 5000 and 0.40 for students who are ranked between 5,001 and 20,000. Although each individual 5,001-20,000 student is not very likely to move into the top 5000 in 11th grade, together they will account for about 23 percent of the year-$t+1$ grade 11 top 5000. If the dynamics were gender-independent, then the fraction of girls in this moving-up group would be close to 40 percent, and this would substantially bring up the average percent female variable in the top 5000.

The next three rows of the Table report the separate 9th to 10th, 10th to 11th, and 11th to 12th grade decompositions that went into the average discussed above. Recall that gender gap widened most from 9th grade to 10th grade. The entry effect is relatively more important at this stage and the mechanical effect does less to offset the other sources of female disadvantage. The changes from 10th to 11th grade are very similar to the overall

[29]The latter matters here because students outside the top 5000 will need to improve by substantially more than the average amount to move into the top 5000.

average. In the 11th to 12th grade transition, the entry effect becomes quite unimportant, but the growth effect is even more important and dropout also plays a role.

The final two rows of the Table focus on more extreme high achievers. Recall that the fraction female in the top 500 declined from 18 percent in 9th grade to just 12 percent in 12th grade. This 35 percent decrease was larger than the 29 percent decrease at the top 5000 level, although it is smaller in percentage point terms (about 2 percentage points per year). The importance of the growth process to the evolution in the gender gap comes through even more strongly here – differences in the probabilities with which boys and girls at each lower rank are able to move into the top 500 are much more important than the other differences we've identified. The entry and dropout effects are both just minor factors, consistent with the view that few true entrants will make it all the way to the top 500 and few students will drop out after earning such high scores.

The bottom row looks at even more extreme high achievers who scored in the top 50 in their grade. Here, the dropout and entry effects continue to fade to insignificance relative to the large growth effect. What remains are the large growth effect and a continuation effect, again offset in large part by the mechanical effect.[30]

The small numbers that come up when doing top 50 calculations may make it easier to understand why the mechanical effect is so large. Going back to the transition matrix illustrated in Figure 2 and inverting the relationship there to count moves into the top 50, we can infer that on average 18.1 of the year-$t+1$ top 50 students will be repeats from the year t top 50, and they will be joined by 8.1 students who ranked between 51 and 100, 6.4 who ranked between 101 and 200, and 5 who ranked from 201-500. On average about 9 percent of the 9th-11th grade top 50 is female. Ranks 51-100 are about 12 percent female. Ranks 101-200 are about 14 percent female. And ranks 201-500 are about 18 percent female. Together, the students moving up from these three lower groups make up about 40 percent of the year-$t+1$ top 50. If they were randomly drawn from their rank groups, then about 16 percent of them would female. Hence, their presence would increase the overall percent female in the top 50 by about $0.4 \times (16-9) \approx 3$ percentage points. The magnitude of these mechanical effects can make our finding of a broadening gender gap even more striking – the widening of the gender gap over the course of high school occurs despite the fact that every year there are many more girls in the set of students well positioned to move into the

[30] In order to account for noise in the decomposition exercise introduced by the local linear regressions, we performed a nonparametric bootstrap of the decomposition procedure, resampling at the student level and holding ranks fixed across 2,000 bootstrap draws. As shown in Table A2 in the Appendix, all terms in Table 6 are estimated with a great deal of precision, with the exception of several factors at the top 50 level.

top 50 (or 500 or 5000) than currently in the top 50 (or 500 or 5000).

5 Reactions to Disappointment

So far, we have tried to improve our understanding of the widening of the gender gap in high math achievement over the high school years by providing detailed descriptive evidence on the dynamics of performance that any potential explanation would have to account for. In this Section, we exploit the multistage nature of the AMC series to provide evidence with a more causal flavor on one potential mechanism: gender differences in how students react to disappointment.

The AMC 10/12 contests are the first stage of a series. Students who score highly enough on the AMC 10/12 are invited to participate in the American Invitational Mathematics Exam (AIME). High AIME scorers are invited to take the USA Math Olympiad (USAMO). High USAMO scorers are invited to the Math Olympiad Summer Program (MOP). Six MOP students are selected to represent the United States at the International Math Olympiad (IMO). Although all but six of the 200,000 plus annual participants in the AMC 10/12 contests eventually fail to reach the next stage, a number of awards are given out along the way and students take pride in how far they advance. For most of the high-achieving students in our sample, the most salient potential accomplishment is qualifying for the AIME. In a typical year in 1999-2006, roughly 500-750 9th graders, 1000-2000 10th graders, 3000-5000 11th graders, and 4000-6000 12th graders qualify. Many who qualify for the AIME will regard this as making their AMC season a success. Qualifying keeps their math competition season alive for another month, and they will list it on their college applications. Many who fall just short of the cutoff for AIME qualification will be disappointed. Our discussions of "reactions to disappointment" should be understood as a shorthand for how students react to this disappointment relative to how they react to the positive feedback that comes with qualifying.

The "rational" response to falling just short might be to redouble one's efforts: not having previously been an AIME qualifier should raise the incremental benefit that qualifying provides to one's resume; and students have learned (given how much students typically improve from year to year) that they have a good chance of qualifying in the subsequent year. However, it also seems plausible that many students might react to the disappointment by investing less in their math skills, whether for behavioral or other rational reasons. In light of the literature on gender differences in self-confidence and interest in competition (e.g., Niederle and Vesterlund 2007; Croson and Gneezy 2009), one could easily imagine

that there are gender differences in this respect.

The rules for advancement from the AMC 10/12 to the AIME are a bit complicated. Students qualify if they score at least 120 on the AMC 10 or 100 on the AMC 12. They also qualify if they are among the top 1 percent of US test takers on the particular (A or B) AMC 10 that they took, or among the top 5 percent on the particular AMC 12. The rules are an ex ante attempt to treat the tests roughly equally, but in practice the ex post level of correctly-measured performance at which the cutoff falls varies from test to test. From our perspective, this is fortuitous in that it makes the AIME qualification "treatment" less collinear with performance.[31] As an initial look at the data, Figure 5 provides an RD-style plot of the probability with which students with scores in each one-point score band cannot be found in the next year's data. Students in the zero band and all students to the right qualified for the AIME. We report the means separately for boys and girls and add separately estimated regression lines on each side of the cutoff. The Figure strongly suggests that there is a discontinuous jump in the probability of dropping out of future participation when students fall just short of the AIME cutoff.

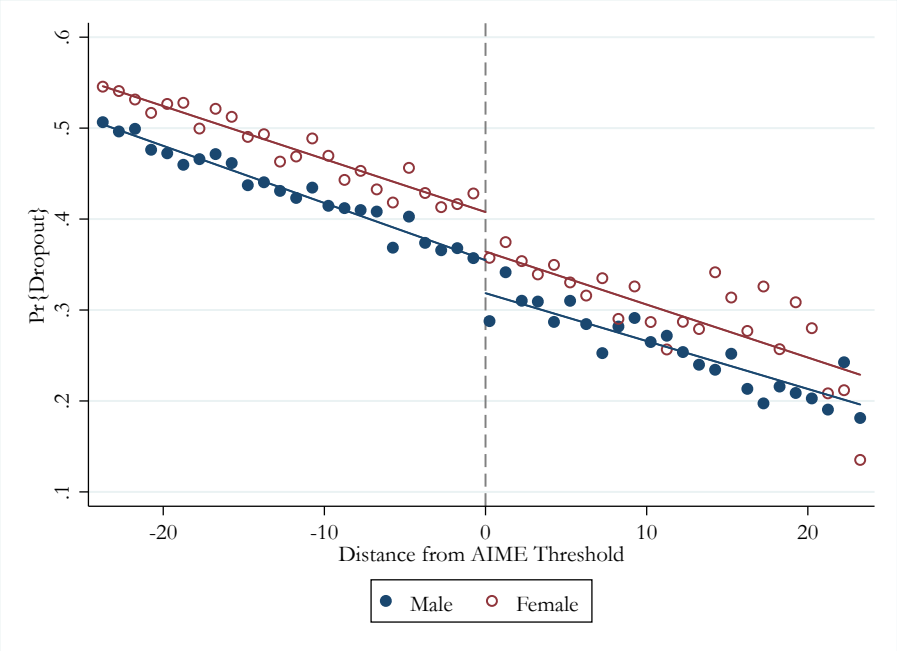
The noisiness of the female data on the right side of the Figure reflects that there are a limited number of girls with scores more than ten points above the AIME cutoff.[32] But in other cases – e.g., the data points for boys exactly at and six points below the AIME cutoff – substantial departures from the regression lines occur despite sample sizes that are quite large. We believe that this reflects the role of unobserved student characteristics that do not covary smoothly with students' scores relative to the cutoff. To illustrate why this is plausible, note that the most common AMC 10 cutoff is 120. In 2002-2006, the unique way to score 120 was to answer all 25 questions and get 20 correct and 5 wrong.[33] The unique way to get the score just below 120, 119.5, was to attempt just 18 of the 25 questions and get 17 correct and 1 wrong, with 7 left blank. The 119.5- and 120-scoring students may therefore be different in unobserved ways; e.g., the 120 students may be quicker, less accurate, and more risk loving. Such discontinuous changes in unobservables of this variety would make the standard RD estimator of the causal effect of AIME qualification

[31] Although some students may be aware that the AIME cutoff for the AMC 10 is often 120, and the cutoff for the AMC 12 is often 100, it would nevertheless be difficult for students to "game" the cutoff and strategically score just above it. If gaming were common, we would expect to see bunching of students right at the cutoff; as shown in Figure A1, the decline in student counts in a neighborhood of the cutoff is smooth for both boys and girls, and there is no evidence of bunching.

[32] See Figure A1 for a histogram, by gender, of the distribution of students relative to the AIME cutoff.

[33] In 1999, the AMC had 30 questions and gave 5 points for a correct answer and 2 for a blank answer. In 2000-2001, the tests gave 6 for a correct answer and 2 for a blank answer. In 2002-2006, the score for a blank answer increased to 2.5.

Figure 5: Reactions to Disappointment – "Dropout" Rates



Notes: Figure reports the raw probability that students with scores around the AIME cutoff "drop out" of participating in the next year, by gender. Running variable is distance between the student's score on the first test he or she took in a given year and the AIME cutoff for that test-year. Lines indicate linear fit within 24 points of cutoff.

inappropriate.

We try to estimate effects in a manner that is robust to this potential problem in two ways. First, we simply estimate a regression similar to our earlier dropout regression, with the addition of a dummy variable for failing to qualify for the AIME, and we restrict our analysis to the subsample of students who were within two correct answers of the AIME cutoff on either side. When the AIME cutoff is 120, we include students who answered 25 questions and got 18 or 19 correct (failing to qualify with a 108 or 114, respectively), as well as students who got 20 or 21 correct (qualifying with a 120 or 126, respectively). In such a sample where the number of students answering each number of questions is roughly balanced, the qualification dummy would be mostly uncorrelated with any function of the number of questions answered, and we would hope that our quadratic in $\log(GradeRank)$ would capture any smooth relationship between higher achievement and dropout rates, whereas the dummy for failing to qualify would capture any discontinuous jump at the year×test-specific cutoff. We also explicitly control for whether a student scored in each

of the subsets of scores, e.g. $\{..., 108, 114, 120, 126, ...\}$ and $\{..., 113.5, 119.5, 125.5\}$ that are possible when students attempt a given number of questions. Given the variety of scoring rules used in different years, this involves adding a total of 18 dummy variables, and can be thought of as an estimator that will give us the causal effect of failing to qualify on the probability of dropping out, provided that the unobservables are smooth across the cutoff once we have controlled for the differences related to the number of questions a student attempted.

The first column of Table 7 presents estimates from this OLS regression.[34] Our main interest in conducting these regressions is on the effect of the disappointing outcome of failing to qualify for the AIME. The main effect on this variable is substantial, 3.7 percentage points, and highly statistically significant. One way to think about the magnitude is that it is comparable to the participation gender gap for 11th grade girls: i.e., it means that an 11th grade boy with a score just below the AIME cutoff will be almost as likely to drop out of participating as an 11th grade girl who scored just above the cutoff.

The second main coefficient of interest is the differential effect that failing to qualify for the AIME has on girls. The estimate indicates that the decrease in the probability of participation is 1.9 percentage points larger for girls than for boys; i.e., girls are even more likely than boys to cease participating in the AMCs when they experience a disappointing outcome (this difference is statistically significant at the 3 percent level). The effect on a girl of just missing the AIME will be the sum of the two estimated coefficients, so girls with scores just below the AIME cutoff will be 5.6 percentage points less likely to participate in the following year than girls who just barely qualify for the AIME. This is consistent with previous literature on gender differences in self-confidence and responses to competition, suggesting that those findings are relevant even to the set of highly accomplished girls we are studying, and providing us with a causal link identifying a factor that contributes to the widening gap. The gap is, however, substantially smaller than the roughly 10 percentage point gap Buser and Yuan (2018) found in a similar analysis of Dutch data, and will only account for a small portion of the observed widening of the gender gap over the high school years.[35]

[34]The regression also includes unreported year and grade fixed effects, a dummy for taking the B-date test, a dummy for taking both the A-date and B-date tests, a quadratic in log(GradeRank), and Female interactions with the linear and quadratic log(GradeRank) terms. We normalize within-grade rank separately within each grade so that the adjusted log of within-grade rank variable has mean zero within each grade for students with scores exactly at the AIME cutoff. With this normalization, for example, the coefficients on the Female $\times$ Grade 9 interaction can be thought of as giving the gender difference in dropout rates for students who qualified for the AIME with the lowest possible score.

[35]The estimate here is sufficiently precise to rule out an effect close to that found in Buser and Yuan

Table 7: Reactions to Disappointment – Regression Discontinuity Evidence on "Dropout" Rates

	Dep. Var.: Dropout $t \to t+1$		
	Sample:		
	Within two answers	Optimal bandwidth	
		Male	Female
	Estimation:		
Variable	OLS	RD with controls	
Female × Grade 9	-0.006		
	(0.011)		
Female × Grade 10	0.009		
	(0.008)		
Female × Grade 11	0.038***		
	(0.006)		
Below AIME Cutoff	0.037***	0.034***	0.042***
	(0.006)	(0.007)	(0.012)
Female × Below AIME Cutoff	0.019*		
	(0.009)		
Bandwidth	10/12	12.043	10.575
Number of observations	139,874	701,686	614,014

Notes: Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Second, we implement an RD estimator with local linear controls for the running variable (distance to AIME cutoff as in Figure 5), an endogenous bandwidth, and robust inference as in Calonico et al. (2014), separately on the male and female samples. In these regressions, we allow for gender-specific nonlinear effects of the running variable on each side of the cutoff and control for year and grade fixed effects, a dummy for taking the B-date test, a dummy for taking both the A-date and B-date tests, and dummies for number of questions attempted.[36]

As reported in the second column of Table 7, the estimated effect of the failing to qualify "treatment" on boys is that it increases their probability of dropout by 3.4 percentage points. This is very similar to the OLS estimate and is similarly significant.[37] The effect

(2018). Their dataset is two order of magnitude smaller, resulting in standard errors that are sufficiently large that they typically cannot rule out a smaller gap of the size we estimate.

[36]In these regressions we use the default optimal bandwidth as implemented in Stata's rdrobust package. An optimal bandwidth of 12 points is selected for males, vs. 10.6 for females.

[37]Both rdrobust results described here are highly significant (p-values ≤ 0.01) according to both conventional and robust inference methods; conventional standard errors are reported in the Table for brevity.

for girls in the third column is somewhat smaller than the OLS estimate at 4.2 percentage points and remains highly significant although the standard error is larger. We can therefore conclude that our finding that both boys and girls are more likely to drop out of future competition is robust to the more flexible allowance for unobserved heterogeneity.[38]

The 0.8 percentage point difference between the male and female dropout effects is estimated sufficiently precisely to rule out the larger difference found in Buser and Yuan (2018), and suggests that this causal channel can only account for a small portion of the observed widening of the gender gap. However, the standard errors are sufficiently large in the "Optimal Bandwidth" specifications that we must also say that whether the gender gap in reactions to disappointment is statistically significant is sensitive to how one controls for potential unobserved heterogeneity.

Disappointment may also affect the performance of students who continue to participate in the AMC tests by affecting the effort students put in over the course of the following year. To look for effects of this type, Table 8 reports coefficient estimates from regressions like those in Table 4 examining the change in within-grade rank between year t and year $t+1$, but using regression discontinuity regressions as in Table 7.

The first main coefficient of interest in this regression is again the coefficient on the dummy for missing the AIME cutoff. In the OLS regression, we get a positive, significant coefficient, which again suggests that students are not responding well to disappointment: students with scores just below the AIME cutoff have a larger increase in their expected year-$t+1$ rank (i.e., they do worse) than do students with scores just above the AIME cutoff. The magnitude is not very large in economic terms – students' ranks are increasing by a little more than 10 percent. However, the fact that it is positive is noteworthy: we have seen that scoring just below the AIME cutoff induces some students to drop out, and the most natural guess would be that these dropouts are relatively weak students, which would result in the pool of continuing students with scores just below the AIME cutoff being positively selected.

In contrast to our earlier result on girls' reacting worse to disappointment in terms of being more likely to drop out, girls who continue participating despite experiencing disappointment show less of a disappointment effect in their performance. This could reflect that the sample of continuing girls is more selected, but could also reflect that girls who do not drop out are less likely to reduce their effort. Regardless, it appears that differences in dropout rates are the channel through which gender differences in reactions

[38] This contrasts with Buser and Yuan (2018)'s being unable to find an effect for boys.

Table 8: Reactions to Disappointment – Regression Discontinuity Evidence on Achievement Gains

	D. V.: $\log(\text{GradeRank}_{t+1}) - \log(\text{GradeRank}_t)$		
	Sample:		
	Within two answers	Optimal bandwith	
		Male	Female
	Estimation:		
Variable	OLS	RD with controls	
Female × Grade 9	0.412***		
	(0.030)		
Female × Grade 10	0.309***		
	(0.022)		
Female × Grade 11	0.285***		
	(0.017)		
Below AIME Cutoff	0.097***	0.048*	0.048
	(0.017)	(0.020)	(0.039)
Female × Below AIME Cutoff	-0.067**		
	(0.025)		
Bandwidth	10/12	13.041	8.156
Number of observations	85,545	324,325	247,507

Notes: Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

to disappointment might contribute to a widening gender gap. The positive coefficient estimate on the Female dummy indicates that (along the lines of what was reported earlier) girls just above and below the AIME cutoff are still improving by less on average than boys with comparable scores.

The remaining columns of the Table provide estimates of the effect of failing to qualify for the AIME on subsequent year performance from the "Optimal Bandwidth" RD procedure. The effect of failing to qualify for the AIME on boys' year-to-year improvement is estimated to be smaller at 0.048 and only marginally significant; the point estimate for girls' reactions is identical at 0.048, though it is not statistically significantly different from zero.

To summarize, students appear to react to the disappointment at falling short of the AIME cutoff both by being more likely to drop out and by improving by less in the subsequent year conditional on not dropping out. The dropout effect may be larger for girls, though this result is sensitive to specification. This could be one factor contributing to the widening of the gender gap, particularly given girls' lower performance in earlier grades and particularly if the observed disappointment effects generalize to other parts of the score

distribution; however, consistent with our above decomposition results, these effects can at most account for a small portion of the observed widening.

6 Conclusions

In this paper, we noted that data from the American Mathematics Competitions indicate that the gender gap among high-achieving math students is already quite large by 9th grade. Girls comprise just 30 percent of the 5000 highest scoring 9th graders on the AMC contests, a figure that is quite close to the 33 percent female representation one sees in the set of high school seniors who earn perfect 800 scores on the mathematics SAT. The AMC tests make it possible to look much farther into the upper tail of mathematics performance and draw consistent distinctions among students whose performance would be top-coded on other tests. Here, we see that the even larger gaps we noted in previous work are already present by 9th grade. Girls comprise just 18 percent of the 500 highest scoring 9th graders on the AMC contests and just 8 percent of the top 50. One of our primary takeaways is that to fully understand the gender gap in high math achievement among high school students, it will be necessary to examine pre-high school data. We hope that our paper will spur further work in this direction.

A second main finding of our paper is that the gender gap in high math achievement widens substantially over the high school years. The largest change occurs between 9th and 10th grades, but it is a fairly steady process clearly visible in every year. The fraction female among students who are among the top 5000 in their grade on the AMC test drops from 30 percent in 9th grade to 22 percent in 12th grade. Among students who are among the top 500 in their grade, the drop is from 18 percent in 9th grade to just 12 percent in 12th grade. These are substantial changes. They would be hard to reconcile with the simplest views of gender gaps stemming from some time-invariant biological difference, and they motivate looking more closely at the year-to-year dynamics of student performance over the high school years.

Our initial analysis of the dynamics of high math achievement brings out several new facts. Two that are particularly important to thinking about the gender gap are that high-achieving students must substantially improve their absolute performance from year to year to maintain their within-cohort rank, and yet within-cohort ranks are still quite persistent. The persistence reinforces our earlier comment that pre-high school factors are important to understanding the gender gap in high school. One can think of the need for substantial improvement to stay in place as deriving from a combination of two effects. One

is that the typical high-achieving math student is substantially improving their knowledge and problem solving skills from year to year. The other is that there are many more students outside the top 500 than in the top 500. Some lower-ranked students are making far-above-average improvements, and this forces highly ranked students to make above-average improvements to maintain their place. The need for these improvements highlights that our high-achieving students are exerting substantial effort in bolstering already highly advanced math skills. There are many, many demands on elite high school students' time that could lead to systematic differences in the opportunity costs of and interest in making such investments.

We have identified four distinct gender-related differences in the dynamics of student performance that contribute to the widening gender gap. In comparison with boys who had the same score in the previous year, high-achieving girls are more likely to drop out of participating in the AMC tests (particularly in 12th grade), and the performance gains of those who do participate again are lower on average and less variable. Girls are also underrepresented in the pool of high-scoring "entrants" whom we could not match to a score in the previous year. Our decompositions point to "growth" differences, the underrepresentation of girls in the set of students who manage to move up from lower ranks to high ranks, as the most important source of the widening gap. The other effects are more moderate in size, but in combination and cumulated over the years also contribute substantially to the observed widening of the gender gap.

In 9th grade, the dearth of female entrants is nearly as important, the effect of high-scoring girls being less able to hold their ranks is consistently part of the story, and from 11th to 12th grade, dropouts become an issue. But the growth differences are consistently the largest effect both across grades and across levels of achievement. In most cases, even by themselves they account for well over 100 percent of the observed broadening of the gender gap. Again, this suggests a line of further inquiry – why are there so few girls who move up substantially relative to their cohort in the later high school years?

Our final Section suggests that reactions to disappointment may be part of the answer. Both boys and girls who experience the disappointing outcome of just barely failing to qualify for the AIME are more likely to not participate in the following year. The dropout effect may be larger for girls, although the significance of this difference is not very clear. Apart from psychological effects related to disappointment, of course, one could also potentially explain such reactions in more standard "rational" ways; e.g., conditional on both boys and girls reacting "irrationally" to disappointment, high-achieving girls might have a

greater breadth of other skills and interests that compete for their time when math seems less promising. We hope to see future work on this as well.

As we have also noted, a limitation of all of our work on the AMC contests is that the data concern performance in a competitive environment. We believe that many of the investments in problem solving skills and mastering precalculus mathematics that the AMC contests measure are investments that will also benefit students in other environments. In our earlier work, we have presented some data on SAT scores consistent with this view, but it would nice to have this complemented with work on the dynamics of achievement as measured with other instruments. It would be even more complementary to be able to track AMC participants forward and examine how participation, achievement, disappointment, etc. on the AMC tests affect outcomes that are well established to be important in later life, e.g. choice of college major, pursuit of postgraduate education, career choices, etc. Agarwal and Gaule (2018) perform such an exercise on a more extreme set of high-achievers in high school math competition – students who advance all the way to the International Mathematics Olympiad – and find that IMO are scores are highly predictive of math publications and citations twenty years in the future, with effects being so strong at the extremes that IMO gold medalists are *fifty* times more likely to win the Fields medal than are graduates of a top 10 math PhD program who did not participate or advance quite this far in high school math contests. AMC scores/participation are surely not as strongly predictive as this of any subsequent achievement, but it would be very interesting to see how they are related to longer-run outcomes.

References

Agarwal, R. and Gaule, P. (2018). Invisible geniuses: Could the knowledge frontier advance faster? IZA Discussion Paper 11977.

Azmat, G., Calsamiglia, C., and Iriberri, N. (2016). Gender differences in response to big stakes. *Journal of the European Economic Association*, 14(6).

Azmat, G. and Petrongolo, B. (2014). Gender and the labor market: What have we learned from field and lab experiments? *Labour Economics*, 30(32-40).

Bertrand, M., Goldin, C., and Katz, L. F. (2010). Dynamics of the gender gap for young professionals in the financial and corporate sectors. *American Economic Journal: Applied Economics*, 2:228–255.

Blau, F. D. and Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3):789–865.

Buser, T. (2016). The impact of losing in a competition on the willingness to seek further challenges. *Management Science*, 62(12):3439–3449.

Buser, T., Niederle, M., and Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *Quarterly Journal of Economics*, 129(3):1409–1447.

Buser, T. and Yuan, H. (2018). Do women give up competing more easily? Evidence from the lab and the dutch math olympiad. *American Economic Journal: Applied Economics*. Forthcoming.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

Carrell, S. E., Page, M. E., and West, J. E. (2010). Sex and science: How professor gender perpetuates the gender gap. *The Quarterly Journal of Economics*, 125(3):1101–1144.

Ceci, S. J., Ginther, D. K., Kahn, S., and Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest*, 15(3):75–141.

Chachra, D., Chen, H. L., Kilgore, D., and Sheppard, S. (2009). Outside the classroom: Gender differences in extracurricular activities in engineering students. In *Proceedings of the 39th Frontiers of Education Conference 2009*.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–474.

Ellison, G. and Swanson, A. (2010). The gender gap in secondary school mathematics at high achievement levels: Evidence from the american mathematics competitions. *Journal of Economic Perspectives*, 24(2):109–128.

Ellison, G. and Swanson, A. (2016). Do schools matter for high math achievement? Evidence from the american mathematics competitions. *American Economic Review*, 106(6):1244–1277.

Fryer, R. G. and Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics*, 2(2):210–240.

Gill, D. and Prowse, V. (2014). Gender differences and dynamics in competition: The role of luck. *Quantitative Economics*, 5(2):351–376.

Ginther, D. K. and Kahn, S. (2004). Women in economics: Moving up or falling off the academic career ladder? *Journal of Economic Perspectives*, 18(3):193–214.

Gneezy, U., Niedele, M., and Rustichini, A. (2003). Performance differences in competitive environments. *Quarterly Journal of Economics*, 118(3):1049–1074.

Goldin, C., Katz, L. F., and Kuziemko, I. (2006). The homecoming of american college women: The reversal of the college gender gap. *The Journal of Economic Perspectives*, 20(4):133–156.

Goldin, C., Kerr, S. P., Olivetti, C., and Barth, E. (2017). The expanding gender earnings gap: Evidence from the LEHD-2000 Census. *American Economic Review, Papers and Proceedings*, 107(5):110–114.

Guiso, L., Monte, F., Sapienza, P., and Zingales, L. (2008). Culture, gender, and math. *Science*, 320(5880):1164–65.

Hedges, L. V. and Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science*, 269(5220):41–45.

Hogarth, R. M., Karelaia, N., and Trujillo, C. A. (2012). When should I quit? Gender differences in exiting competitions. *Journal of Economic Behavior and Organization*, 83(1):136–150.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., and Williams, C. C. (2008). Gender similarities characterize math performance. *Science*, 321(5888):494.

Iriberri, N. and ReyBiel, P. (2018). Competitive pressure widens the gender gap in performance: Evidence from a twostage competition in mathematics. *The Economic Journal*.

Joensen, J. S. and Nielsen, H. S. (2016). Mathematics and gender: Heterogeneity in causes and consequences. *The Economic Journal*, 126(593):1129–1163.

Niederle, M., Segal, C., and Vesterlund, L. (2013). How costly is diversity? Affirmative action in light of gender differences in competitiveness. *Management Science*, 59(1):1–16.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics*, 122(3):1067–1101.

Niederle, M. and Vesterlund, L. (2010). Explaining the gender gap in math test scores: The role of competition. *Journal of Economic Perspectives*, 24(2):129–144.

Pope, D. G. and Sydnor, J. R. (2010). Geographic variation in the gender differences in test scores. *Journal of Economic Perspectives*, 24(2):95–108.

Xie, Y. and Shauman, K. A. (2003). *Women in Science: Career Processes and Outcomes*. Cambridge, MA: Harvard University Press.

Appendix

A Score Adjustments

The adjusted scores for students who took a test other than the AMC 12A in year t are given by formulas of the form

$$AdjustedScore_{ijt} = b_{0jt} + b_{1jt}Score_{ijt},$$

where i indexes the student and $j \in \{10A, 10B, 12A\}$ indexes the test that the student took in year t. For each t from 2000 to 2006, the construction of year t adjusted scores is based on a regression of year $t+1$ AMC 12 scores on dummies for the test taken in year t, interactions between these dummies and the score on the year t test, and a dummy for whether the year $t+1$ score was on the 12B.[39] For each year t test, this regression gives the predicted year $t+1$ AMC 12A score as an affine function of the year t score. We define the adjusted score for each year t test as the score on the year t AMC 12A that has the same predicted year $t+1$ score given the regression estimates.

We cannot adjust 2007 scores in the same way because our dataset does not contain 2008 scores. For the AMC 10 tests, we instead set the slope coefficient $b_{1j2007} \equiv \frac{1}{5}\sum_{t=2002}^{2006} b_{1jt}$ equal to the average of the b_{1jt} for the previous five years, and set $b_{0j2007} = \frac{1}{5}\sum_{j=2002}^{2006} b_{0jt} + \Delta_j$, which is an average of the constants from the previous five years plus an adjustment factor that reflects whether each 2007 test appears to be easier or harder relative to the 2007 AMC 12A than were previous-year AMC 10's relative to their contemporaneous AMC 12A's, in light of data on students who took both tests in each year.[40] To determine the adjustments, we run regressions examining the difference between B-date scores and A-date scores for students who took both tests in each year on dummies for which tests they took,

$$ScoreB_{it} - ScoreA_{it} = c_{10A,t}Dummy10A_{it} - c_{10B,t}Dummy10B_{it} - c_{12B,t}Dummy12B_{it} + \epsilon_{it},$$

and set $\Delta_j = c_{j,2007} - \frac{1}{5}\sum_{t=2002}^{2006} c_{j,t}$. We adjust AMC 12B scores in a somewhat similar manner, but imposing that $b_{1,12B,2007} = 1$ rather than estimating the coefficient.[41] We then set $b_{0,12B,2007} = \frac{1}{5}\sum_{j=2002}^{2006} b_{0,12B,t} + c_{12B,2007} - \frac{1}{5}\sum_{t=2002}^{2006} c_{12B,t} + \left(\frac{1}{5}\sum_{j=2002}^{2006} b_{1,12B,t} - 1\right)\overline{X}$ where $\overline{X} \approx 99.4$ is the mean AMC 12B score among students who scored at least 90 on the 2007 AMC 12 and attend a school that did not offer the 2007 AMC 10A.[42]

To give a feel for the linear adjustments, Table A1 reports the contemporaneous AMC

[39] We run these regressions on the set of students who scored at least 90 on their year t AMC 12 or at least 105 on the AMC 10 because we will be primarily interested in high-achieving students. The linear functional forms appear to fit well for students with scores in this range.

[40] We do not use this approach in all years because the population of students taking A and B date tests are quite different and students may select into taking B date test in addition to an A date test if their A date score was below what they expected.

[41] If we had estimated the coefficient via the same procedure we use for the AMC 10 tests, the estimated coefficient would have been 1.015.

[42] The final term is a small correction designed to correct for the fact that we are imposing a slope coefficient of one when the regression coefficients on the AMC 12B score are not exactly equal to one in the regressions used to estimate the b_{012Bt}.

Table A1: Adjusted Scores for AMC 10A, 10B, and 12B Scores of 100 and 150

	AMC 12A equivalents					
	AMC 10A		AMC 10B		AMC 12B	
Year	100	150	100	150	100	150
2000	83.9	116.7				
2001	64.7	116.1				
2002	90.4	135.2	83.4	131.0	89.0	143.8
2003	92.1	133.2	86.6	126.1	101.9	139.3
2004	94.1	141.0	82.9	131.6	94.4	149.5
2005	91.9	134.7	87.2	133.6	97.6	158.3
2006	94.4	135.5	96.7	131.6	99.5	146.2
2007	79.2	122.6	86.0	129.4	95.9	146.6

12A scores corresponding to scores of 100 and 150 on each of the other tests. Recall that 100 is roughly the 95th percentile score on the AMC 12 and 150 is a perfect score. The left and center parts of the Table give the AMC 10A-to-AMC 12 and AMC10B-to-AMC 12 conversions. The median AMC 10 test will have its scores adjusted downward by 13 points at the 100 level and by 18 points at the 150 level. There is some variation around this – the AMC 10 seems to have been much easier in its first two years and the 2006 AMC 10 tests appear to have been nearly as hard as the 2006 AMC 12 for students scoring around 100 – but most tests are within one question (6 points) of the average relative difficulty level. Most of the AMC 12B adjustments are also less than the point value of one question.

B Variance Calculation

To provide some estimates of the relative importance of measurement error and true performance increases we compare performance changes over multiple years. Suppose that year-t performance $y_{it} = a_{it} + \epsilon_{it}$ reflects both student i's true ability a_{it} and an additive mean-zero measurement error ϵ_{it}. Suppose that ability evolves according to $a_{it+1} = \alpha_0 + \alpha_1 a_{it} + u_{it+1}$. And suppose that the measurement errors ϵ_{it} are independent of all other terms. We then have

$$\begin{aligned} \mathrm{Var}(y_{it+1} - \alpha_1 y_{it}) &= \mathrm{Var}(\epsilon_{it+1}) + \alpha_1^2 \mathrm{Var}(\epsilon_{it}) + \mathrm{Var}(u_{it}) \\ \mathrm{Cov}(y_{it+1} - \alpha_1 y_{it}, y_{it} - \alpha_1 y_{it-1})) &= -\alpha_1 \mathrm{Var}(\epsilon_{it}) + \mathrm{Cov}(u_{it+1}, u_{it}) \end{aligned}$$

It is natural to assume that the true improvements u_{it} are positively correlated, as some students are presumably working harder on improving than others. In this case, the covariance term as provides a lower bound on the measurement error variance:

$$\mathrm{Var}(\epsilon_{it}) \geq -\frac{1}{\alpha_1} \mathrm{Cov}(y_{it+1} - \alpha_1 y_{it}, y_{it} - \alpha_1 y_{it-1}),$$

The lower bound will be close to the true value if $Cov(u_{it+1}, u_{it})$ is small. If we use $\log(Rank_{it})$ as the performance measure y_{it}, estimate α_1 via an IV regression of y_{it+1} on y_{it} using y_{it-1} as an instrument run on the sample of students who took the AMC for three consecutive years, and compute the above variances and covariances on the same sample, we get a lower bound estimate of $\text{Var}(\epsilon_{it})$ that corresponds to ϵ_{it} having a standard deviation of 0.62. This indicates that a substantial portion of the apparent year-to-year variation in scores is due to the measurement error of the test as a measure of underlying achievement.

If we assume that this lower bound also applies to $\text{Var}(\epsilon_{it+1})$, we can also plug into the formulas above to get an upper bound on the standard deviation of u_{it}, which describes the heterogeneity in students' true improvement from year to year. Again, this should be close to the true value if the covariance of year-to-year improvement is low. This estimate corresponds to u_{it} having a standard deviation of 0.37, which suggests that there is substantial heterogeneity in students' true improvement from year to year, albeit not nearly as much as naively looking at year-to-year changes in scores might suggest.

C Decomposition Appendix

As discussed in the main text, our analysis focuses on changes in the fraction μ^f_{Xt} of students in achievement group X at time t who are female. For example, the group X might be the top 5000 scorers. We relate this to various aspects of differences in the boys' and girls' transition matrices. To define these, we write μ^f_{rt} for the fraction female at rank r at time t. We define a_{rX} as the fraction of students at rank r at time t who achieve a score in X at time $t+1$. We will define this both for numerical rank groups $r \in X$ and for the set of students who do not participate at time $t+1$, which we denote by $r = NP$. Write a^f_{rX} and a^f_{rNP} for the analogous objects for female students. Write N_{rt} and N_X for the number of students at rank r at time t and the number with ranks in the set X. Write μ^{new}_{Xt+1} for the fraction of students in group X at time $t+1$ who had not participated at time t and $\mu^{f,\text{new}}_{Xt+1}$ for the fraction of students in group X at time $t+1$ who are female and who had not participated at t. Finally, define each of the component Δ factors in Proposition 1 as follows:

$$
\begin{aligned}
\Delta^{\text{drop}}_X &= \frac{1}{N_X} \sum_{r \neq NP} \mu^f_{rt}(a_{rNP} - a^f_{rNP}) \frac{a^f_{rX}}{1 - a^f_{rNP}} N_{rt} \\
\Delta^{\text{cont}}_X &= \frac{1}{N_X} \sum_{r \in X} \mu^f_{rt}(1 - a_{rNP}) \left(\frac{a^f_{rX}}{1 - a^f_{rNP}} - \frac{a_{rX}}{1 - a_{rNP}} \right) N_{rt} \\
\Delta^{\text{grow}}_X &= \frac{1}{N_X} \sum_{r \notin X, r \neq NP} \mu^f_{rt}(1 - a_{rNP}) \left(\frac{a^f_{rX}}{1 - a^f_{rNP}} - \frac{a_{rX}}{1 - a_{rNP}} \right) N_{rt} \\
\Delta^{\text{entry}}_X &= \mu^{f,\text{new}}_{Xt+1} - \mu^f_{Xt} \mu^{\text{new}}_{Xt+1} \\
\Delta^{\text{mech}}_X &= \frac{1}{N_X} \sum_{r \neq NP} \mu^f_{rt} a_{rX} N_{rt} - \mu^f_{Xt}(1 - \mu^{\text{new}}_{Xt+1})
\end{aligned}
$$

Proof of Proposition 1 Proposition 1 states that the change in the fraction female in group X can be written as

$$\mu^f_{Xt+1} - \mu^f_{Xt} = \Delta^{\text{drop}}_X + \Delta^{\text{cont}}_X + \Delta^{\text{grow}}_X + \Delta^{\text{entry}}_X + \Delta^{\text{mech}}_X.$$

With all transition probabilities like a_{rX} representing the realized fraction of students at rank r at time t who score in the top X at $t+1$, and taking the sum over all ranks that

have at least one female student at time t, we have:

$$\begin{aligned}
\mu^f_{Xt+1} - \mu^f_{Xt} &= \frac{1}{N_X}\sum_{r\neq NP}\mu^f_{rt}a^f_{rX}N_{rt} + \mu^{f,\text{new}}_{Xt+1} - \mu^f_{Xt} \\
&= \frac{1}{N_X}\sum_{r\neq NP}\mu^f_{rt}a_{rX}N_{rt} + \mu^{f,\text{new}}_{Xt+1} - \mu^f_{Xt} + \frac{1}{N_X}\sum_{r\neq NP}\mu^f_{rt}(a^f_{rX} - a_{rX})N_{rt} \\
&= \frac{1}{N_X}\sum_{r\neq NP}\mu^f_{rt}a_{rX}N_{rt} - \mu^f_{Xt}(1-\mu^{\text{new}}_{Xt+1}) + \mu^{f,\text{new}}_{Xt+1} - \mu^f_{Xt}\mu^{\text{new}}_{Xt+1} \\
&\quad + \frac{1}{N_X}\sum_{r\neq NP}\mu^f_{rt}(a^f_{rX} - a_{rX})N_{rt} \\
&= \Delta^{\text{mech}}_X + \Delta^{\text{entry}}_X + \frac{1}{N_X}\sum_{r\neq NP}\mu^f_{rt}(1-a_{rNP})\left(\frac{a^f_{rX}}{1-a_{rNP}} - \frac{a_{rX}}{1-a_{rNP}}\right)N_{rt} \\
&= \Delta^{\text{mech}}_X + \Delta^{\text{entry}}_X + \frac{1}{N_X}\sum_{r\neq NP}\mu^f_{rt}(1-a_{rNP})\left(\frac{a^f_{rX}}{1-a^f_{rNP}} - \frac{a_{rX}}{1-a_{rNP}}\right)N_{rt} \\
&\quad + \frac{1}{N_X}\sum_{r\neq NP}\mu^f_{rt}(1-a_{rNP})\left(\frac{a^f_{rX}}{1-a_{rNP}} - \frac{a^f_{rX}}{1-a^f_{rNP}}\right)N_{rt} \\
&= \Delta^{\text{mech}}_X + \Delta^{\text{entry}}_X + \Delta^{\text{cont}}_X + \Delta^{\text{grow}}_X + \\
&\quad + \frac{1}{N_X}\sum_{r\neq NP}\mu^f_{rt}\left((1-a^f_{rNP})\frac{a^f_{rX}}{1-a^f_{rNP}} - (1-a_{rNP})\frac{a^f_{rX}}{1-a^f_{rNP}}\right)N_{rt} \\
&= \Delta^{\text{mech}}_X + \Delta^{\text{entry}}_X + \Delta^{\text{cont}}_X + \Delta^{\text{grow}}_X + \frac{1}{N_X}\sum_{r\neq NP}\mu^f_{rt}(a_{rNP} - a^f_{rNP})\left(\frac{a^f_{rX}}{1-a^f_{rNP}}\right)N_{rt} \\
&= \Delta^{\text{mech}}_X + \Delta^{\text{entry}}_X + \Delta^{\text{cont}}_X + \Delta^{\text{grow}}_X + \Delta^{\text{drop}}_X \qquad \square
\end{aligned}$$

The decomposition in Proposition 1 is an accounting identity that will hold exactly in the data for any one year if one defines the top X so that it has exactly X students in each year, sets all transition probabilities like a_{rX} to be the actual fraction of students at rank r at t who scored in the top X at $t+1$, and is consistent in what one plugs in for the multiple occurrences of conditional transition probabilities like $a^f_{rX}/(1-a^f_{rNP})$ that are undefined because the denominator is zero.[43] It will also hold exactly in data from multiple years if one uses appropriate weighted averages.

We instead implement the decomposition by estimating the transition probabilities both for the full population and for girls as smooth functions of the initial year rank via local linear regressions with $\log(Rank)$ as the right-hand-side variable.[44] This makes all of the

[43] Plugging in different numbers will, however, alter the results of the decomposition, shifting between attributing changes to dropouts and to attributing them to differential growth and continuation rates.

[44] For example, to estimate a_{rX} for X being the top 500 students, we use a dependent variable which is one for all students who score strictly in the top 500, zero for all students who are outside the top 500 (including students who did not take the test), and an intermediate fraction for all students whose $t+1$ score is at the level that spans the boundary. The fraction female in the top X similarly uses fractional

transition probabilities continuous in rank and provides a natural definition for the conditional probabilities, avoiding any indeterminacies. We do this separately for students in 9th, 10th, and 11th grades, pooling the data for all six cohorts within each regression.

Table A2: Decomposition of Declines in Fraction Female, with Confidence Intervals

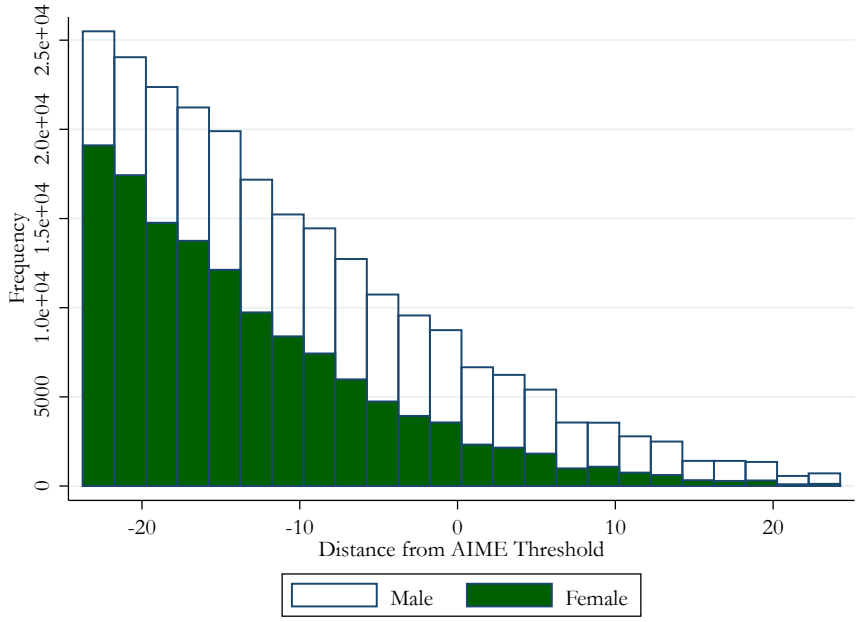
Grade Level	Achievement Level	Change in % Female	Decomposition of decline				
			Drop	Cont	Grow	Entry	Mech
Average	Top 5000	-3.1	-0.4	-1.2	-3.6	-1.1	3.5
		[-3.2,-2.9]	[-0.5,-0.4]	[-1.3,-1.1]	[-3.8,-3.5]	[-1.2,-1.0]	[3.4,3.6]
$9 \rightarrow 10$	Top 5000	-4.6	-0.1	-1.4	-2.8	-2.2	1.5
		[-5.1,-4.2]	[-0.2,0.0]	[-1.5,-1.2]	[-3.1,-2.7]	[-2.5,-1.9]	[1.9,2.2]
$10 \rightarrow 11$	Top 5000	-2.3	-0.4	-1.1	-3.7	-0.8	3.6
		[-2.7,-1.9]	[-0.5,-0.3]	[-1.2,-1.0]	[-3.9,-3.5]	[-1.0,-0.5]	[3.7,4.0]
$11 \rightarrow 12$	Top 5000	-2.1	-0.9	-1.1	-4.3	-0.3	4.7
		[-2.7,-1.9]	[-1.0,-0.7]	[-1.3,-1.0]	[-4.5,-4.1]	[-0.5,-0.2]	[4.5,4.9]
Average	Top 500	-2.0	-0.3	-0.9	-4.5	-0.4	3.8
		[-2.4,-1.5]	[-0.4,-0.1]	[-1.2,-0.7]	[-4.8,-4.1]	[-0.7,-0.1]	[3.9,4.5]
Average	Top 50	-0.5	-0.3	-0.8	-3.0	0.2	3.2
		[-1.6,0.5]	[-0.6,0.3]	[-1.7,0.1]	[-4.1,-2.2]	[-0.4,0.7]	[2.4,4.2]

Notes: Table reports 90 percent confidence intervals from nonparametric bootstrap, resampling at the student level and holding ranks fixed across 2,000 draws.

counting for students at the score spanning the boundary.

D Regression Discontinuity Appendix

Figure A1: Regression Discontinuity Support – Histogram by Gender



Notes: Figure reports the count of male and female students in each relative score bin (the regression discontinuity running variable). Running variable is distance between the student's score on the first test he or she took in a given year and the AIME cutoff for that test-year.