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

Glenn Ellison
MIT and NBER

and

Ashley Swanson
University of Pennsylvania and NBER

October 2018

[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 in high math achievement over the high school years using data from 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 dropout rates and in the mean and variance of year-to-year improvements among continuing students. 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 is 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 the American Mathematics Competitions (AMC) to examine the dynamics of the gender gap over the high school years within a 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.

In previous work, we argued that two features of the AMC contests make them a unique and valuable instrument for studying high-achieving math students: the AMC tests are much better than the standard tests at identifying and distinguishing among very high-achieving students; and the tests are taken by many of the very best math students in

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

[2] See 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.

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. These attractive features 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 with our previous work, this is relevant to interpreting our finding about the 9th grade gender gap being large. It should have less of an effect on the dynamic analyses, which compare boys and girls who initially selected into competition and scored similarly, that are the focus of most of this paper. That girls and boys may later drop out of participating in the AMC contests can even be seen as a fourth advantage: 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 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

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

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. Niederle and Vesterlund (2007), for example, show in a laboratory experiment in which men and women have equal ability that men are more likely than women to opt for tournament-based compensation, which suggests that gender differences in preferences over competition could affect entry.[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

[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.

likely to drop out of participating. And girls are underrepresented among the high-scoring entrants. To clarify the relative importance of the various diferences 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, and that the tendency to drop out after experiencing disappointment is more common among girls.

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

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 analysis of students' reaction to disappointment can be seen as providing real world evidence of 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 Section 4 of Buser and Yuan (2018), which 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. Our much larger 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 percentage points for boys (a 9 percent effect, relative to dropout rates among boys scoring just above the cutoff) and 5 percentage points for girls (a 15 percent effect), with standard errors below 1 percentage point. Thus, we confirm 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.[9] 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).

[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] 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.

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 U.S. high 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.[10] A second is that the tests are very popular among the very best math students in the US.[11] 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

[10] 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.

[11] 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 presidental scholar candidates attend schools that offer the AMC. Relevant to assessing participation among extreme high achievers, they note that at least 80% 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% of the students at participating schools with 800s on the SAT math take the AMC contests.

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 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.[12]

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.

[12] 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.

In addition to the scores, the records contain a school identifier, the state in which the school is located, an anonymization of the 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.[13]

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 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.[14]

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

[13] 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 $grade_t - grade_{t'} = t - t'$.

[14] 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.

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.[15] 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 participation 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 takeway from the participation levels is that high-achieving girls and boys are roughly equally likely to participate in 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.

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

[15] 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.

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

Table 1: Average annual AMC participation and scores by gender and grade level

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 and mixes in more students from lower percentiles with students who would truly be in the highest percentile if repeatedly tested. But the magnitude of the difference does suggest that there are some gender-related differences in participation rates, which would have to take the form of female participation rates being a bit lower among the very highest scorers on the SAT and slightly higher among students with somewhat lower SAT scores. The former difference would be expected given the literature on gender differences in attitudes toward competition. The latter would need to reflect some other gender-related difference, such as being more likely to comply with a teacher's suggestion to take a test even when the student will not do well.

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 students on the AMC.[16] 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 it widens substantially over the course of high school.

Table 2 reports the fraction of AMC high scorers in each grade who are female for various definitions of high scoring. The first row examines 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. In the upper left cell, we see that there is a substantial gender gap in 9th grade: only 30.5 percent of the high-scoring 9th graders are female.[17] Comparing from left-to-right across the columns, we see that the gender gap widens in each subsequent year. By 12th grade, only 21.8 percent of the 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.

Level of achievement	% female among top students in grade				% change
	Grade 9	Grade 10	Grade 11	Grade 12	$9 \rightarrow 12$
Top 5000	30.5	25.8	24.3	21.8	-29%
Top 1000	21.1	17.6	16.3	14.6	-31%
Top 500	17.9	15.5	13.1	11.7	-35%
Top 100	11.0	11.4	8.0	7.5	-32%
Top 50	8.4	8.4	7.7	6.8	-19%

Table 2: Percent female by grade and achievement level

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

[17] The figure reports the average across the six cohorts that we observe for all four of their high school years of the fraction female within each cohort.

Subsequent rows present comparable figures using more and more demanding definitions of high-achieving, going all the way out to a definition that allows only the top 1 percent of our initial high-achieving pool. Looking from left-to-right within each row, we see 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 fraction female from 9th grade to 12th is between 19 percent and 35 percent in every row. The proportional decline over the course of high school appears largest when we use the top 500 definition and smallest when we use the top 50 definition, but it should be kept in mind that the estimates in the final row will be quite noisy given the small sample sizes.

Ellison and Swanson (2010) highlighted that the gender gap is much larger when one examines more extreme high achievers. The first column of Table 2 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. Indeed, the decline from the top row to the bottom row is roughly similar in each column. 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 in high school. What is clearly an issue that must be understood by studying the high school years, however, is that the gender gap is widening 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.[18] To give some sense of how performance grows over time, Table 3 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.

Within-grade rank	Corresponding overall rank				Decrease in overall rank to maintain rank in grade		
	Grade 9	Grade 10	Grade 11	Grade 12	$9 \to 10$	$10 \to 11$	$11 \to 12$
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%

Table 3: Growth in absolute performance: 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

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

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 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.[19]

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 3 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 4 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

[19] 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.

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 to maintain a given within-grade rank in Table 3 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.

	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	

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

Table 4: IV regressions of growth in absolute performance as a function of initial performance relative to cohort

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. A comparison of performance changes over multiple years can give some insight on the relative magnitudes of measurement error and true performance increases. 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

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

One would assume that the true improvements u_{it} are positively correlated, as some students are presumably working harder on improving than others. Hence, one could think of the covariance term as providing a lower bound on the measurement error variance:

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

that would 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.

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. Table 5 presents an estimated rank-to-rank transition matrix. For example the element in the third column of the first row presents the probability that a student who is among the top 50 in their cohort in year t will rank from 101 to 200 in year $t+1$.[20]

[20]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.

Year t rank	Probability of reaching year $t+1$ rank group conditional on year t rank							
	1-50	51-100	101-200	201-500	501-1000	1001-5000	5000+	No match
1-50	0.36	0.14	0.13	0.12	0.06	0.04	0.01	0.14
51-100	0.16	0.12	0.13	0.19	0.10	0.11	0.04	0.15
101-200	0.06	0.08	0.11	0.17	0.12	0.20	0.04	0.20
201-500	0.02	0.03	0.06	0.15	0.14	0.28	0.11	0.21
501-1000	0.01	0.01	0.03	0.09	0.10	0.33	0.19	0.25
1001-5000	0.00	0.00	0.00	0.02	0.04	0.23	0.36	0.35

Table 5: Persistence in math performance: forward transition matrix of 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 Table 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 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 additional rows. 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 substantial improvements are quite unlikely. Four group improvements are close to zero percent events – the probability of moving from the 501-1000 group into the top 50 is estimated to be slightly above 0.5 percent. However, making less sizable gains is also quite uncommon. Only 6-16 percent of the student in each rank group make enough of an improvement to move into a higher rank group in the next year. And only 2-6 percent make enough of an improvement to move up two or more groups. Declines in within-cohort ranks are much more common. For example, 53 percent of students ranked from 201-500 in year t will score outside the top 500 in year $t+1$.

A third observation is that dropping out of participation is relevant even among high-achieving students. 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.[21]

One final comment on the Table 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 Table 5 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. Table 6 reports in each column the fraction of students who achieved the rank corresponding to that column in 12th grade who were in each rank category in 9th grade. Note that the numbers in each column now sum to one. At the very highest levels of achievement, the performance persistence we noted earlier remains striking. There are more holdovers from the 9th-grade top 50 in the 12th-grade top 50 than there are students who have moved up from the entire 201-40,000 range. Only 5 percent scored outside the top 1000 as 9th graders. Although there are a substantial fraction, 35 percent, whom we were unable to

[21] 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 mechanical match, which is about 18 percent on average across grades. Several factors are involved in the superiority of this manual match over our mechanical 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 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.

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 severe when we are matching across a three year span.

Within-grade rank in grade 9	Fraction of students in each grade 12 rank group who were in each rank group in grade 9					
	Within-grade rank in grade 12					
	1-50	51-100	101-200	201-500	501-1000	1001-5000
1-50	0.25	0.11	0.06	0.02	0.01	0.00
51-100	0.10	0.10	0.06	0.03	0.01	0.00
101-200	0.09	0.07	0.07	0.05	0.02	0.00
201-500	0.10	0.11	0.12	0.09	0.05	0.02
501-1000	0.06	0.09	0.07	0.10	0.07	0.03
1001-5000	0.03	0.09	0.14	0.18	0.20	0.15
5001+	0.02	0.01	0.05	0.07	0.12	0.19
No match	0.35	0.42	0.43	0.47	0.53	0.62

Table 6: Early performance of top math students: backward transition matrix of probability of each within-9th-grade rank group for students in each within-12th-grade rank group

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 had not done as well in 9th grade, with 18 percent coming from the 1001-5000 range and 7 percent from the above 5000 range.

At the lower (but still high) levels of 12th grade achievement in the Table, 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 7 presents estimates from an OLS regression which has the change in each student's within-grade rank as the dependent variable. 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.[22] The

[22]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 nonmissing. The regressions also include unreported year and grade dummies. The $\log(\text{GradeRank}_t)$ control is adjusted by subtracting the sample mean.

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.

A primary coefficient of interest in the regression is the coefficient on the Female dummy. It is highly significant and indicates that girls are improving by less from year to year than boys by about 30 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 part 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.

	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	0.30***	0.32***
	(0.010)	(0.043)
Adj $\log(\text{GradeRank}_t)$	-0.18***	-0.08***
	(0.004)	(0.016)
Female $\times$ Adj $\log(\text{GradeRank}_t)$	-0.04***	0.02
	(0.011)	(0.051)
$\hat{\sigma}^2_{\text{male}}$	1.53***	2.21***
	(0.009)	(0.034)
$\hat{\sigma}^2_{\text{female}}$	1.22***	1.99***
	(0.015)	(0.079)
Number of observations	81,570	9,682

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

Table 7: Gender difference in within-cohort rank dynamics for high-achieving students

In the above regression, there is 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 7 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. As an additional robustness

test, 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} for whether each year-t high scorer could not be found in the year-$t+1$ data, and regress this on the student's within-grade rank in year t, the square of this variable, a Female dummy, a Female $\times$ rank interaction, and grade, year and B-date dummies.[23]

The first column of Table 8 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 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.

	Dep. Var.: Dropout $t \rightarrow t+1$				
	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***	0.009	0.017**	0.046***	0.007
	(0.003)	(0.005)	(0.005)	(0.006)	(0.010)
Adj log(GradeRank_t)	0.071***	0.077***	0.066***	0.071***	0.033***
	(0.002)	(0.004)	(0.004)	(0.004)	(0.006)
Adj log(GradeRank_t)^2	0.008***	0.009***	0.008***	0.008***	0.005*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Female $\times$ Adj log(GradeRank_t)	0.001	0.001	0.003	-0.003	-0.012
	(0.004)	(0.006)	(0.006)	(0.007)	(0.012)
Number of observations	119,325	39,284	39,747	40,294	12,020

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

Table 8: Gender difference in within-cohort rank dynamics for high-achieving students

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

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 4.6 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.[24] 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 8 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 $t+1$. The point estimate of the gender-related difference in dropout rates is also much smaller, just 0.7 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. 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, Table 9 lists 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.[25] 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 in the

[24] 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.

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

percent female is about one percentage point. 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. The lower number of female students among students who achieve high scores after not having taken the AMC in the previous year is a fourth contributor to the broadening of the gender gap over the high school years.

	Fraction of students in each rank group who are female						
	1-50	51-100	101-200	201-500	501-1000	1001-5000	Average
All year-t students	0.09	0.12	0.14	0.18	0.21	0.29	0.17
Year-$t+1$ entrants	0.10	0.11	0.14	0.16	0.20	0.26	0.16

Table 9: Gender composition of highly ranked students new to the AMC in comparison with highly ranked students from the previous year

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.

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.) 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.

Proposition 1 *The change in the fraction female in group X can be written as*[26]

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

where

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

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^{entry}_X, 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^{mech}_X, 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.

[26] See Appendix B for proof.

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.[27] The sign of the mechanical effect Δ_X^{mech} will depend on which is larger.

The decomposition 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 $a_{rX}^f/(1-a_{rNP}^f)$ that are undefined because the denominator is zero.[28] 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.[29] This makes all of the 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.

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 10 presents a decomposition of this change.[30] It indicates that

[27] We can make this intuition precise by further decomposing Δ_X^{mech} into terms reflecting these two effects. Write μ_{Xt+1}^{cont} for the fraction of students in the top X in year $t+1$ who had been in the top X in year t and μ_{Xt+1}^{grow} for the fraction of students in the top X in year $t+1$ who participated but were not in the top X in year t. We then have $\Delta_X^{\text{mech}} = \Delta_X^{\text{mech,cont}} + \Delta_X^{\text{mech,grow}}$ where $\Delta_X^{\text{mech,cont}} = \frac{1}{N_X}\sum_{r\in X}\mu_{rt}^f a_{rX} N_{rt} - \mu_{Xt}^f \mu_{Xt+1}^{\text{cont}}$ and $\Delta_X^{\text{mech,grow}} = \frac{1}{N_X}\sum_{r\notin X, r\neq NP}\mu_{rt}^f a_{rX} N_{rt} - \mu_{Xt}^f \mu_{Xt+1}^{\text{grow}}$. We would expect $\Delta_X^{\text{mech,cont}}$ to be negative and $\Delta_X^{\text{mech,grow}}$ to be positive.

[28] 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.

[29] 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 counting for students at the score spanning the boundary.

[30] 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.

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 r 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 variance of students' improvements.[31]

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.

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

Table 10: Decomposition of declines in fraction female

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 0.263 for students in the top 5000 and 0.401 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

[31] 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.

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 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 r 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.[32]

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 Table 5 and inverting the relationship there to count moves into the top 50, we can infer that 18.1 of the year-$t+1$ top 50 will be repeats from last year's top 50, but they will be joined by 8.1 students who ranked between 51 and 100 last year, 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

[32]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 10 are estimated with a great deal of precision, with the exception of several factors at the top 50 level.

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 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 on which we have focused are the first stage of a multistage series. Students who score highly enough on the AMC 10 or 12 are invited to participate in the American Invitational Mathematics Exam (AIME). Roughly 500 high scorers from that exam advance to the USA Math Olympiad (USAMO). A few dozen high scorers on that test are invited to the Math Olympiad Summer Program (MOP). And in the end, six MOP students are selected to represent the United States at the International Math Olympiad. While this technically means that all but six of the 200,000 plus students who participate in the AMC 10/12 contests each year eventually lose, 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 make it will regard AIME qualification as making their AMC season a success. It keeps their math competition season alive for another month, and they will plan to list it on their college applications. Many who fall just short of the cutoff for AIME qualification will know that they have fallen just short and be quite disappointed. Our discussions of "reactions to disappointment" should be understood as a shorthand for the reaction to the difference between the disappointment of failing to qualify and the excitement and positive feedback

that comes with qualifying.

The "rational" response to falling just short would probably be to redouble one's efforts: not having been an AIME qualifier should raise the incremental benefit that qualifying provides to one's resume; students have gotten a signal that with a little more preparation they would have made it; and they've also learned (given how much students typically improve from year to year) that they have a good chance of qualifying in the subsequent year. But in practice, it seems likely that some students will instead be discouraged and decide to invest less in their math skills. 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 will vary depending on which test a student took: when the absolute score thresholds bind it is better to have taken whichever test date (A or B) is easier; and the percentage thresholds will not exactly correct for differences between the talent in the AMC 10 and AMC 12 student pools and do not attempt to correct for the A vs. B distinction, even though the B pool is historically stronger.

One's initial thought might be that this provides a classic opportunity to apply a standard regression-discontinuity design: each year-10/12-A/B combination provides a cutoff; students with scores just above and below the cutoff are similar in performance but get different disappointment treatments. Due to the details of the scoring, however, a standard RD design is not really appropriate. For an RD design, one wants the relationship between unobserved student characteristics and the score to be smooth around the cutoff. The vagaries of how the AMC scores the exams and penalizes guessing make this unlikely to be a good assumption. For example, the AMC 10 cutoff is most often 120. The unique way to get a 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, 1 wrong, with 7 left blank. It is easy to imagine that the 119.5 and 120-scoring students are different in unobserved ways. The 120 students may be quicker, less accurate,

[33]The AMC contests changed the guessing penalty after 2007, so the calculations here are different from what they would be on current contests.

and more risk loving. A more subtle observation is that the 119.5 students might also be less sophisticated in not having realized that their chances of reaching a potential 120 cutoff would have been higher if they had guessed on two additional questions.[34] Differences of this variety are not smooth or even monotone. The score just above 120, 120.5, was obtained by answering 20 of the 25 questions and having 18 correct and 2 wrong, with 5 left blank. Students with this outcome may be more like 119.5 students than 120 students on the speed/accuracy/risk-loving dimension, but different from both (although more like 120 students) on the strategic sophistication dimension.

The structure of the AMC scoring is, however, very well suited to applying a variant of the RD design. The patterns in the score vs. "number of questions attempted" relationship repeat every six points (which corresponds to the point value of a correct vs. an incorrect answer.) If one compares students with scores that range from 6 points below the AIME cutoff to 5.5 points above, one can think of the comparison as a pooled comparison of students in 12 different score pairs, each of which should be similar in unobserved dimensions. For example, students who score 120 (20 correct, 5 wrong) should be similar in unobservables to students who score 114 (19 correct, 6 wrong). And students who score 119.5 (17 correct, 1 wrong) should be similar to students who score 125.5 (18 correct, 0 wrong). Within each pair, students who make one fewer mistake will be a little stronger on average, but it seems the difference should not be very large and can presumably be controlled for reasonably well, as in our earlier regressions which used the log of a student's within-grade rank as a flexible control for the general tendency of higher performing students to be less likely to drop out of participating in the AMCs. Given that the location of the cutoff varies from year to year and from test to test within each year, we can hope to identify (and hence control for) this smooth relationship fairly accurately even in a dataset restricted to students who are within 6 points of the AIME cutoff. Given such a control, a regression including a dummy for whether a student was just above or below the AIME cutoff on the test they took can be thought of as investigating the effect of a "disappointment" treatment that is close to randomly assigned.

Table 11 presents estimates from regressions which use the above approach to examine the effect of disappointment on the probability that students will drop out of participating in the AMC tests. The first column contains estimates from a linear regression like those which we used previously to examine gender-related differences in dropout rates, but estimated on

[34] If all 18 of the answers they gave were correct, their score would exceed 120 even if the two added guesses were both incorrect. If one of the 18 was wrong, they would surpass 120 if either of the two added guesses were correct, but not if they made no guesses.

the sample of students in the close-to-the-cutoff window described above and adding dummy variables for failing to qualify for the AIME. In light of the differences noted earlier, we also allow the coefficient on Female dummy to vary across grades.[35] We again normalize within-grade rank separately within each grade, but here have done the normalization 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, the coefficients on the Female × Grade X interactions can be thought of as giving the gender difference in dropout rates for students who qualified for the AIME with the lowest possible score. Here, we find that 11th grade girls are significantly more likely to drop out of participation (by 4.1 percentage points) even after having qualified for the AIME. The corresponding estimates for 9th and 10th grade girls are smaller and not statistically significant, but the standard errors are such that we also cannot rule out effects on the order of 2-3 percentage points.

	Dep. Var.: Dropout $t \to t+1$			
	Sample: within 1 question of AIME cutoff			
Variable	All grades	Grade 9	Grade 10	Grade 11
Female × Grade 9	0.004	-0.037		
	(0.015)	(0.022)		
Female × Grade 10	0.013		0.014	
	(0.010)		(0.014)	
Female × Grade 11	0.041***			0.047***
	(0.007)			(0.009)
Below AIME Cutoff	0.029***	0.011	0.033**	0.035***
	(0.008)	(0.017)	(0.011)	(0.008)
Female × Below AIME Cutoff	0.022*	0.071*	0.018	0.018
	(0.010)	(0.032)	(0.020)	(0.012)
Adj log(GradeRank$_t$)	0.030***	0.026	0.025*	0.066***
	(0.009)	(0.022)	(0.012)	(0.012)
Adj log(GradeRank$_t$)^2	0.015**	0.012	0.017	0.009
	(0.005)	(0.015)	(0.010)	(0.007)
Female × Adj log(GradeRank$_t$)	-0.017	0.003	-0.007	-0.024*
	(0.009)	(0.027)	(0.016)	(0.011)
Number of observations	71,853	6,455	17,017	48,381

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

Table 11: Reaction to disappointment: dropout rates around the AIME cutoff

Our main interest in conducting these regressions was on the effect of the disappointing

[35] As before, the regressions also include unreported year, grade, and B-test dummies.

outcome of failing to qualify for the AIME. The main effect on this variable is substantial, 2.9 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 2.2 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. 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 5.1 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. It is also noteworthy in connection with our earlier finding that the differential dynamic that contributes most to the widening gender gap is that we see fewer girls moving up and substantially improving their within-cohort rank from year to year.

The final three columns of the Table report coefficients from regressions estimated separately on the 9th, 10th, and 11th grade subsamples. Note that the 11th grade subsample is much larger than the others: there are many more 11th graders than younger students with scores very close to the AIME cutoff. Several results are clearly significant just in the 11th grade subsample: girls who just barely qualify for the AIME are more likely to drop than boys with the same score; and students of either gender who just miss out on qualifying are more likely to cease participating. The coefficient on the male-female difference in the effect of disappointment is also similar to the full sample estimate, but now significant only at the 14 percent level. While one might have hoped that the regressions in the other columns would shed light on whether disappointment effects and gender differences therein are different for 9th and 10th graders, the smaller sample sizes prevent us from saying much. The point estimates on both the main effect and the female interaction are positive in all grades and none are significantly different from the all-grades estimates. The main effect is significantly different from zero at the 1 percent level in the 10th grade subsample and the female interaction is significant at the 3 percent level in the 9th grade sample, but mostly we are unable to provide much in the way of statistically significant conclusions

when examining the earlier grades in isolation. Note that we also lose much of our ability to identify the general relationship between performance levels and dropout rates, which is a critical control in these regressions.

	D. V.: $\log(\text{GradeRank}_{t+1}) - \log(\text{GradeRank}_t)$			
	Sample: within 1 question of AIME cutoff			
Variable	All grades	Grade 9	Grade 10	Grade 11
Female	0.303***	0.343***	0.287***	0.296***
	(0.019)	(0.076)	(0.042)	(0.023)
Below AIME Cutoff	0.106***	0.241***	0.173***	0.079***
	(0.017)	(0.059)	(0.032)	(0.023)
Female × Below AIME Cutoff	-0.074*	-0.047	-0.031	-0.104**
	(0.029)	(0.115)	(0.062)	(0.034)
Adj $\log(\text{GradeRank}_t)$	-0.422***	-0.463***	-0.457***	-0.436***
	(0.018)	(0.056)	(0.031)	(0.027)
Female × Adj $\log(\text{GradeRank}_t)$	0.022	0.060	0.026	0.035
	(0.026)	(0.098)	(0.052)	(0.032)
Number of observations	45,761	4,821	12,140	28,800

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

Table 12: Reaction to disappointment: changes in performance subsequent to scoring around the AIME cutoff

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 12 reports coefficient estimates from regressions like those in Table 7 examining the change in within-grade rank between year t and year $t+1$, but as in Table 11 restricting the sample to students with year t scores close to the AIME cutoff and including dummies for missing the AIME.[36]

The first main coefficient of interest in this regression is again the coefficient on the dummy for missing an AIME. 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

[36]The regressions also include unreported grade and year dummies, and a dummy for the year t test being a B-date test.

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 main channel through which gender differences in reactions to disappointment 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 final three columns of the Table provide estimates from regressions run separately on the 9th, 10th, and 11th grade subsamples. The estimates on the control variables are quite consistent across the grades. The main effect of falling below the AIME cutoff is noticeably larger in 9th and 10th grades than in 11th. Recall that the disappointment-related dropout effects were less precisely estimated (and perhaps smaller) in the younger grades. The significance here means that we can now conclude that adverse effects of disappointment are visible in all three grades. The interaction between the female dummy and falling below the AIME cutoff is significant only in the 11th grade sample.

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 is larger for girls and will be one factor contributing to the widening of the gender gap.

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 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. And we note that the dropout effect is even larger for girls. 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.

References

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

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.

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.

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.[37] 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.[38] 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.[39] 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.[40]

To give a feel for the linear adjustments, Table A1 reports the contemporaneous AMC 12A scores corresponding to scores of 100 and 150 on each of the other tests. Recall that

[37] 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.

[38] 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.

[39] 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.

[40] 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}.

	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

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

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 Decomposition Appendix

Proof of Proposition 1

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 we have

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

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]

Table A2: Decomposition of declines in fraction female, with 90% confidence intervals from nonparametric bootstrap