Do Schools Matter for High Math Achievement? Evidence from the American Mathematics Competitions

Glenn Ellison and Ashley Swanson[1]

November 19, 2015

[1]Ellison: Department of Economics, MIT, 77 Massachusetts Avenue, Building E18, Room 269F, Cambridge, MA 02139, and NBER (e-mail: gellison@mit.edu); Swanson: The Wharton School, University of Pennsylvania, 3641 Locust Walk, CPC 302, Philadelphia, PA 19104, and NBER (e-mail: aswans@wharton.upenn.edu). 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. Hongkai Zhang and Sicong Shen provided outstanding research assistance. Victor Chernozhukov provided important ideas and help with the methodology. We thank the editor and several anonymous referees for their comments; we are particularly grateful to an anonymous referee who generously provided us with data that allowed us to extend our analysis. We also thank David Card and Jesse Rothstein for help with data matching. Financial support was provided by the Sloan Foundation and the Toulouse Network for Information Technology. Much of the work was carried out while the first author was a Visiting Researcher at Microsoft Research. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

Abstract

This paper uses data from the American Mathematics Competitions to examine the rates at which different high schools produce high-achieving math students. There are large differences in the frequency with which students from seemingly similar schools reach high achievement levels. The distribution of unexplained school effects includes a thick tail of schools that produce many more high-achieving students than is typical. Several additional analyses suggest that the differences are not primarily due to unobserved differences in student characteristics. The differences are persistent across time, suggesting that differences in the effectiveness of educational programs are not primarily due to direct peer effects.

Keywords: gifted education, unobserved heterogeneity, school quality, semiparametric count data models, AMC, American Mathematics Competition, mathematics education

1 Introduction

High-achieving students make important contributions to scientific and technical fields, and educational productivity has been heralded as a vital source of comparative advantage for the United States.[1] It is therefore troubling that the U.S. trails most OECD countries not only in average math performance, but also in the fraction of students that earn very high math scores.[2] It is yet unclear how best the U.S. might address this significant shortcoming. It seems obvious that high-quality schools would play an important role, but several recent papers that examine gifted programs and elite magnet schools have found further troubling evidence that these schools/programs do not appear to benefit marginal students.[3] In this paper, we examine the questions of whether schools matter for high math achievement and whether there are many more students in the U.S who would have reached high math achievement levels in a different environment. We examine the rates at which different schools produce high scorers in the American Mathematics Competition (AMC) contests and note that there are substantial differences across seemingly similar schools. We conduct several additional analyses to investigate whether these differences may be due to unobserved differences in underlying student ability. The analysis suggests that there are substantial differences in the effectiveness of schools' educational programs and that the high-achieving students we observe are a small subset of those who would have reached high math achievement levels in a different environment.

The primary data source for our study is the Mathematical Association of America's AMC 12 contest. The contest is a 25-question multiple choice test on precalculus topics given annually to over 100,000 U.S. students at about 3,000 high schools. The primary advantages of the AMC are that the test is explicitly designed to test depth of knowledge and

[1] Two notable examples of high-achieving high school students with a large economic impact are Microsoft's Bill Gates, who coauthored a computer science paper as a Harvard freshman, and Google's Sergey Brin, who finished in the top 55 on the 1992 Putnam Exam. See Hoxby (2002) for a discussion of education as a source of U.S. comparative advantage in human-capital intensive industries, Krueger and Lindahl (2001) and Hanushek and Woessmann (2008) for surveys of the empirical literature on education and growth, and Altonji (1995), Levine and Zimmerman (1995), Rose and Betts (2004), Joensen and Nielsen (2009), and Altonji, Blom, and Meghir (2012) for evidence on math and the labor market.

[2] Hanushek, Peterson, and Woessmann (2011) note that "most of the world's industrialized nations" have a higher percentage of students reaching advanced levels on the 2006 PISA (Programme for International Student Assessment) test than does the U.S. On the 2009 PISA math test, just 1.9% of U.S. students achieved "Level 6" scores, whereas the OECD average was 3.1% and Singapore had 15.6% of its students at this level. PISA's reading tests indicate that the U.S. is good at producing students with very high verbal achievement: the U.S.'s percentage of "Level 6" reading students is well above the OECD average (1.5% vs. 0.8%).

[3] See Abdulkadiroglu, Angrist, and Pathak (2014), Bui, Craig, and Imberman (2014), and Dobbie and Fryer (2014).

high-level problem-solving skills and can distinguish among students at very high achievement levels. The primary drawback, which will influence how we conduct the analysis, is that the test is taken by a non-random self-selected sample of students. We examine different levels of "high" achievement. Many of our analyses focus on counts of students within a school who score at least 100 on the 2007 AMC 12. One can think of this as roughly comparable in difficulty to scoring 800 math SAT, but measured using a more reliable test that emphasizes greater depth of knowledge.[4] We also examine students at a substantially higher achievement level: those scoring at least 120 on the AMC 12, which puts them well above the 99.9^{th} percentile in the US population.[5] We match AMC schools to data from the census, National Center for Education Statistics (NCES), College Board, and ACT to obtain demographic data and other covariates and conduct most of our analyses on the subsample of public, coed, non-magnet, non-charter U.S. high schools that administer the AMC 12 and could be matched to the other databases. Section 2 discusses the data in more detail.

Section 3 begins with our most basic observation: some schools produce many more AMC high scorers than do other schools with similar demographics. Negative binomial regressions show that there are a number of strong demographic predictors of high achievement. For example, a one percentage point increase in the fraction of adults with graduate degrees is associated with a seven percent increase in the number of AMC high scorers, and a one percentage point increase in the fraction of the population that is Asian-American is associated with a two percent increase. But the most important finding is that the demographic effects are far from sufficient to account for the observed differences in the rates at which different schools are producing high-achieving students. The excess variance parameter in the negative binomial model can be thought of as an estimate of the magnitude of the unobserved multiplicative "school effects" that would be necessary to account for the observed dispersion. We find that the necessary variance is 0.73 at the 100-AMC level and 2.18 at the 120-AMC level; e.g., it is as if a school that is one standard deviation above average produces students who score at least 100 on the AMC 12 at a rate that is 85% ($\approx \sqrt{0.73}$) greater than average.

Section 4 examines the distribution of the unobserved "school effects." The motivation for examining these distributions, rather than being satisfied with knowing the variance, is

[4] 800 is the highest possible score on the math SAT and is achieved by approximately one percent of SAT-takers.

[5] One noteworthy student who reached this level five years earlier is Facebook's Mark Zuckerberg, whose name appears on the AMC's 2002 distinguished honor roll for having scored 121.5.

similar to that for the literature on the heterogeneity in teacher value-added: it is useful to know, for example, if the variance seems to be due to the existence of a subset of low-performing schools in which students are very unlikely to become AMC high scorers, or if it is due to a small (or big) set of schools that produce high-scorers at much (or slightly) higher than average rate.[6] Counts of high-achieving students are inherently small, so one cannot precisely estimate a school fixed effect for any one school. But one can estimate the distribution of school effects across schools.[7] Formally, we implement a nonparametric estimator for the distribution of the unobserved component in a model in which schools produce high scorers at Poisson rates which differ due to observed school and local area characteristics and an unobserved component.[8] We estimate that many schools are producing high scorers at a well below average rate, e.g., 32% of schools appear to be producing high scorers at less than one-half of the average rate. And a striking finding is that there appears to be thick upper tail of schools that produce AMC high scorers at many times the average rate. For example, we estimate that more than 11% of schools are producing AMC high scorers at more than twice the rate one would expect given their demographics and 1% of schools are producing AMC high scorers at more than five times the average rate.

The "school effects" we estimate can be thought of as indexes that conflate multiple factors that lead to heterogeneity in outcomes. They will reflect differences in causal effects across school environments. But they will also reflect other less interesting sources of outcome heterogeneity: differences due to potentially observable demographic differences not captured by variables in our dataset; unobservable differences in student ability due to location decisions of parents of gifted children; and even less interestingly, differences in the fraction of the high-achieving math students at each school who take the AMC 12 test. Section 5 presents several additional analyses aimed at assessing whether effects of the less interesting types are likely to account for a substantial portion of the heterogeneity we have found. To examine the effects of heterogeneity in participation rates, we compare the magnitudes at the AMC 100 and AMC 120 levels. (We argue that selection into test-taking is not important at the higher level.) We explore the importance of unobserved heterogene-

[6] See, for example, Gordon, Kane, and Staiger (2006) for plots of the estimated distribution of teacher qualities obtained by shrinking estimated teacher fixed effects to take out purely random variation.

[7] Although we have AMC data at the individual level, we only observe some coarse demographic variables and hence have chosen to estimate school effects rather than, for example, multilevel models as in Skrondal and Rabe-Hesketh (2009).

[8] The method involves a series expansion similar to that of Gurmu, Rilstone, and Stern (1999) but relying on a different characterization of the likelihoods designed to be more appropriate for potentially fat-tailed distributions. Appendix I contains more detail on the estimation along with Monte Carlo estimates illustrating the performance of the estimator.

ity in student populations in two ways: we examine counts of students achieving perfect scores on the math sections of the SAT and ACT which should be similarly affected by demographic differences but less sensitive to differences in the depth of knowledge developed by the school environment; and we compare school effects estimated from counts of all high-scoring students with school effects estimated from counts of high-scoring girls. A motivation here is that unobserved demographic differences that impact location decisions should not differ much between high-ability male and female students, whereas differences in school environments, such as whether a school's high-level math programs are female-friendly, would lead to gender-related differences.[9] Taken together, these analyses provide evidence that unobserved heterogeneity in performance across schools is much greater than can be explained by student heterogeneity or selection into taking the test.

Section 6 presents several analyses intended to provide insight into the mechanisms that may be making some schools more effective than others. We begin by exploring one very simple mechanism: there could be large effects on the number of students with upper-tail scores if some schools increase their students' scores by a constant and the distribution of scores has a thin-tailed distribution. We look for such a mechanism by including school-average SAT/ACT scores as a covariate. We find little effect, suggesting that our school effects reflect something salient for high-achieving students in particular. Although we have discussed our school effects as indexes reflecting heterogeneity in outcomes, it is more accurate to think of them as measures of the strength of the forces needed to produce the observed degree of clustering of high-scoring students. Such clustering could be generated by unobserved heterogeneity in school quality, but it could also be an artifact of strong peer effects among high-scoring students. Using a formal model similar in spirit to that of Ellison and Glaeser (1997), we note that, with a single observation on each school, one cannot distinguish unobserved heterogeneity in school quality from peer effects as a source of such clustering. However, one can estimate the relative importance of school quality vs. peer effects with multiple observations per school if school quality is persistent and peer effects are not felt across periods. We present such an estimate derived from comparing the agglomeration of 2007 high scorers to the coagglomeration of 2007 and 2003 high scorers. It suggests that differences in school environments are not primarily due to within-cohort peer effects. Finally, we present some qualitative observations on some of the most unexpectedly successful schools. Among our observations are that a number of these schools have a long-serving "star" teacher.

[9] Ellison and Swanson (2010) document that, at the highest performance levels, female high-scorers are drawn from a much smaller set of super-elite schools than are male high-scorers.

Our work is related to a number of literatures. One is the literature on the effectiveness of elite schools and gifted programs. Recent papers by Abdulkadiroglu, Angrist, and Pathak (2014), Dobbie and Fryer (2014), and Bui, Craig, and Imberman (2014) examine the effects of elite magnet schools and gifted programs on high ability students using regression discontinuity designs. Their common finding that the programs have no impact on the test scores of marginal admitted students is striking given that one might have expected the students to benefit from superior peers even absent superior instruction and has attracted a great deal of attention. Our contrasting suggestion that schools are important for gifted students could potentially be reconciled in various ways: the tests we study assess more in-depth understanding and problem-solving skills; effects of programs on marginal admitted students could be very different from the effects on students in the opposite tail; or it could be that the particular gifted programs studied in the previous papers are not very effective but the programs of many other schools in our sample are.[10]

A second related literature on gifted students is that on low-income students with high SAT/ACT scores missing from the student bodies of elite colleges. Several authors have noted that there are many such students and a proximate cause is that many low income high school students with high SAT/ACT scores do not apply to elite colleges.[11] Hoxby and Avery (2013) provide the most comprehensive analysis and document that such students are disproportionately found in areas where they are unlikely to have the opportunity to attend a selective high school, to study with teachers who attended selective colleges, and to interact with many high-achieving peers. Our findings are somewhat analogous in that we are suggesting that there are many students who could have achieved an educational distinction in a different environment. One difference, however, is that (by focusing on schools that offer the AMC) we are focusing on missing students from within a set of relatively high-achieving high schools. Presumably, the set of students identified in the papers noted above would be a substantial additional source of students who could have been high math achievers.[12]

There is a much larger literature on quality differences across schools affecting average achievement. This includes many papers examining how inputs affect achievement and

[10] Angrist and Rokkanen (2012) examine additional test scores of admitted students and argue that Boston Latin School also appears to have little impact on the performance of students farther from the cutoff on state proficiency tests.

[11] See Bowen, *et al.* (2005), Avery, *et al.* (2006), and Pallais and Turner (2006).

[12] Two other relevant literatures on high-achieving students are the literature on cross-sectional differences in high achievement (e.g., Hanushek, Peterson, and Woessman (2011), Pope and Sydnor (2010), and Andreescu, *et al.* (2008)), and the literature on how proficiency-focused reforms may harm high-achieving students (e.g., Krieg (2008), Neal and Schanzenbach (2010), and Dee and Jacob (2009).)

papers that focus on differences in productivity related to competition, vouchers, charter schools, etc.[13] Many of these papers control for selection effects and estimate causal effects relevant to school reform debates. The smaller literature examining residual variance is more closely related.[14] These papers generally find that unobserved school-level heterogeneity is much less important than are student and neighborhood characteristics for predicting test scores, graduation rates, and labor market outcomes. Our findings that schools appear to matter a great deal to high-achieving students may sound conflicting, but could be reconciled in several ways: it has been noted previously that differences that are small relative to within-school variation and/or demographic differences can still be large in absolute terms; environmental differences that produce small differences in mean scores could have a magnified impact when one looks at tail outcomes; or there could be more true heterogeneity in school quality relevant to high math achievement.

Our methodology is related to the literature on measuring agglomeration, including papers such as Ellison and Glaeser (1997), Marcon and Puech (2003), Duranton and Overman (2005), Bayer and Timmins (2007), and Ellison, Glaeser, and Kerr (2012). Whereas many indexes are motivated as a scalar representation of the strength of agglomerative forces, we quantify agglomerative forces by estimating a distribution of effect sizes. Our discussion of peer effects and unobserved heterogeneity is also related to Graham (2008), which derives general conditions under which peer effects and unobserved heterogeneity can be separately identified by looking at residual covariances and includes an application to peer effects in the STAR experiment.

2 The American Mathematics Competitions

The American Mathematics Competitions is the largest and most prestigious series of math competitions for U.S. high schools students. The AMC 12 contest is a 25 question multiple choice test administered in over 3,000 U.S. high schools. About 100,000 students participated in 2007.[15] Our primary motivation for examining AMC data is that we feel that the AMC 12 is superior to any other test administered at comparable scale in its reliability and

[13] See Coleman (1966), Hanushek (1986), Card and Krueger (1992), Hoxby (2000), Hoxby (2000, 2002), Angrist, *et al.* (2002), Hoxby, *et al.* (2009), Abdulkadiroglu, *et al.* (2011), and Dobbie and Fryer (2013).

[14] See Jencks and Brown (1975), Solon, Page, and Duncan (2000), Rothstein (2005), and Altonji and Mansfield (2010).

[15] The AMC 12 is the first stage of a series. In 2007, about 8,000 AMC 12 high scorers were invited to take the AIME. About 500 AMC/AIME high scorers were then invited to take the USAMO. Finally, about fifty high scorers from the USAMO were invited to a summer training program, from which six students were chosen to form the U.S. team for the International Mathematical Olympiad.

validity for identifying students at very high levels of math achievement.

Regarding reliability, the most natural comparison is to the math portion of the SAT reasoning test. We regard the SAT as not reliable above the 97th percentile: when students who score an 800 (a perfect score which is the 99^{th} percentile) retake the math SAT, only 15% score 800 again and their average retake score of 752 is a 97^{th} percentile score. Scoring 100 on the AMC 12 can be thought of as roughly comparable in difficulty to scoring 800 on the SAT. Scoring 120 on the AMC 12 can be thought of as at least an order of magnitude more difficult and places students above the 99.9^{th} percentile of the SAT population. In a striking contrast to the SAT, the AMC 12 remains well calibrated even at the higher of these levels.[16]

Regarding validity, we would argue first that a casual inspection of the tests strongly suggests that the AMC 12 is superior to the SAT as a test of important math skills. A student is awarded an 800 score on the math SAT only if he or she can work through 54 relatively straightforward problems in 70 minutes without making a single mistake. The AMC 12 consists of 25 problems on a wide range of precalculus topics. They generally require greater depth of knowledge and problem solving skills than SAT questions. To score 100 on the AMC 12, a student need only solve 14 of the 25 problems in 75 minutes. To give a sense of what this entails, Figure 3 in the Appendix contains questions 13 through 20 from the 2007 AMC 12. Questions are arranged in increasing difficulty, so students will probably need to solve at least two or three of these problems to score 100. Scoring 120 on the AMC 12 is much harder. It requires answering at least 19 questions correctly. Hence, it requires that students be able to work much more quickly and solve essentially all of the questions shown in the Figure.

Statistical evidence of validity can also be provided by examining how high AMC scorers do when taking other tests. To develop evidence on how well AMC scores predict success on SAT-style tests, we obtained AMC and SAT scores for 195 quasi-randomly selected MIT applicants. Table 1 reports the mean score and the fraction of 800's that students obtained when they first took the math SAT as a function of their 2007 (11th grade) AMC 12 score.[17] The Table provides striking evidence that the AMC 12 is much more powerful predictor of students' ability to achieve a high SAT score than is the SAT itself. To develop

[16]The AMC 12 is offered on two different dates each year. As noted in Ellison and Swanson (2010), students who scored 95 to 105 on the first 2007 date and retook the test averaged 103 with a standard deviation of 11 on the retake. Students who scored 115 to 125 on the first date averaged 119 with a standard deviation of 10 on the retake.

[17]The final column contains data from the College Board on how students with an 800 SAT perform when they retake it.

evidence on how well AMC scores predict success in a very different important environment – solving more difficult open-ended problems – we gathered data from two states, Georgia and Massachusetts, which have state-level competition series that start with a broadly administered multiple choice test and culminate with tests that ask students to solve more open-ended problems and write some formal proofs.[18] Georgia's 2007 contest started with 1,440 students from 88 schools. At the end, three students were named as winners and 27 were awarded Honorable Mention. Despite these very long odds and the proof orientation of the final test, 15 of the 20 students at participating schools who had scored at least 120 on the 2007 AMC 12 made the top 30.[19] The 2007 Massachusetts contest started with 2,000+ students from roughly 60 schools and 20 students were named as winners. Students scoring at least 120 on the 2007 AMC 12 could not possibly have succeeded at a comparable rate on the Massachustts contest – there were 47 such students at participating schools. But the slightly more select sample of 22 students who had scored at least 129 on the AMC 12 were again remarkably successful given the long odds and very strong field: 12 of the 22 finished in the top 20. We conclude that the AMC 12 does remarkably well in identifying students with high-level math skills.

The primary drawbacks of the AMC 12 as a research tool are that it is only administered in nonrandom subset of U.S. high schools and that students self-select into participation. The 3,000 AMC-offering schools are only about 10% of the total number of U.S. high schools. The fraction of high-achieving U.S. high school students who can take the AMC 12 in their high school is much higher than 10% – schools' decisions to offer the AMC 12 are highly correlated with student demographics – but still probably only about 50%.[20] The lack of universal administration makes it impossible for us to provide estimates of the number of high-achieving math students nationwide. However, we can work around this drawback by focusing our analyses on heterogeneity in achievement within the set nonmagnet, noncharter, coeducational public schools that administer the AMC 12.

The second drawback of the AMC is more problematic: some schools may be more aggressive than others in encouraging their best math students to take the AMC 12. This will be a source of apparent performance differences even when educational environments are identical. The main thing we will do to try to get a sense for how this may be affecting our results is to perform some analyses both on the set of students scoring 100 on the AMC

[18] See Figure 4 in the Appendix for two sample problems from Georgia's 2007 contest.

[19] Each of the three winners scored at least 132 on the AMC 12.

[20] One statistic we can provide to support this is that 58% of the students who were named Presidential Scholar candidates (an honor based on having high combined math plus reading SAT or ACT scores) attended a high school that offered the AMC 12.

12 and on the set of students scoring 120 on the AMC 12.

3 Data

The 2007 AMC 12 contest consisted of two separate tests administered on different dates: schools could give the "12A" exam on February 6, 2007 and/or the "12B" exam on February 21, 2007.[21] Our raw data are at the individual level data and contain the test date (A or B), score, student ID, school ID, grade, gender, and home ZIP code. For most of our analyses we work with school-level aggregates. We merge the AMC data with several other databases. First, we match the schools to the schools in the NCES Common Core of Data (CCD) and Private School Survey (PSS) by school name, city, and state.[22] Second, we obtain additional demographic variables by matching the schools to census data on the ZIP code in which the school is located.[23] Finally, we matched the schools to a database containing counts of SAT/ACT takers, average SAT/ACT scores and counts of students with perfect scores by school.[24]

Our primary variables of interest will be counts of students in each school scoring at least 100 or 120 on the AMC 12.[25] In defining these variables we use the count of students scoring at least the cutoff on the 12A contest if the school administered the 12A and the count scoring at least the cutoff on the 12B exam if the school offered only the 12B.[26] In

[21] The primary motivation for having to test dates is to facilitate the participation of schools that may be on vacation or have some other conflict with one date. About 64% of U.S. schools administer only the 12A exam, with 28% administering only the the 12B, and 8% administering both.

[22] The AMC school IDs are usually a school's CEEB code. We obtain the school name, city, and state using the CEEB search program on the College Board's website. Of the 3,730 schools with numerical CEEB codes in the AMC data, 3,105 were matched to schools in the NCES data. 311 of the unmatched 625 schools do not appear in the NCES data because they are not in the U.S. A further 160 could not be matched because the AMC school IDs were not valid CEEB codes. The remaining 154 unmatched schools will include among others, private schools that do not appear in the NCES survey data because private schools are not required to fill out the PSS. Among the matched schools, a further three were dropped because they were missing covariates used in the estimations described below.

[23] Such data were available for 3,021 of 3,105 AMC schools.

[24] We drop schools from the sample if we are unable to match to the SAT data. We also drop if the ACT data are missing and the school is located in a state where more than 20% of students take the ACT, or if the SAT data are missing and the school is located in a state where more than 20% of students take the SAT. We drop all data from Arkansas, Illinois, and Wyoming, where all ACT data are missing. A total of 157 schools are dropped for one of these reasons.

[25] The AMC 12A and 12B are not necessarily identical in difficulty. To control for this, we adjust the AMC 12B cutoffs such that, within the sample of 2,286 students who took both exams, the count of students scoring above 100 on the 12A is equivalent to the count of students scoring above the adjusted cutoff on the 12B. We perform a similar adjustment for the 120 cutoff. Based on this adjustment, we use 12B cutoffs of 108 and 123 as equivalent to 100 and 120 on the 12A.

[26] Note that we do not count students from a school that offered both the 12A and the 12B if they did not participate in the 12A and then scored above the cutoff on the 12B. We count in this way to be conservative in measuring heterogeneity: schools that administer the AMC 12 on both dates are disproportionately high-

our analyses using SAT/ACT data we use both school-average scores and the number of students with perfect scores. In defining these variables, we use a student's SAT score if the student took the SAT and the ACT score if the student did not.[27]

Most of our analyses will be run on the set of public, coed, nonmagnet, noncharter schools that administered the AMC 12. Eliminating magnet schools is important to make it feasible to control for the quality of the student population using available demographic data on the school and its ZIP code. In addition to dropping 800 schools listed by the NCES as being private, non-coed, magnet, or charter schools, and three schools missing NCES data on school demographics, we dropped an additional 76 schools after a manual examination: we examined all schools that were among the 200 largest positive outliers in preliminary regressions of the count of AMC and SAT high scorers on demographic variables, as well as schools outside the 1^{st} to 99^{th} percentile range in percent of female students, and dropped schools that offered a special program that seemed likely to attract high-achieving math/science students from outside a neighborhood attendance area.

Note that in aggressively dropping schools with magnet-like features, we are omitting many schools with programs explicitly designed to promote high math achievement. Hence, while our inability to eliminate all self-selection is a factor that will lead us to overstate the prevalence of high-achieving schools, our data construction also has a potentially strong bias working in the opposite direction. For example, we drop Rockdale County HS in Conyers, GA from our dataset because it houses a school-within-a-school, the Rockdale Magnet School for Science and Technology, which enrolled about 40 students per grade in 2007. However, even if one were to think of these students as the 40 best in the entirety of Rockdale County, the school's performance would be impressive. The entire county's population is only about 85,000 and given its demographics (e.g. majority free/reduced lunch, 60% African American, 2% Asian American), the three AMC 12 high scorers from the school in 2007 are about 10 times as many as one would have expected to find in the entire county. And the fact that the school has special programs makes it all the more plausible that features of the school environment are responsible for this success.

Table 2 contains summary statistics for the merged database of 1,984 sample schools

achieving schools and we want to eliminate any advantage they may obtain from increasing participation by offering both test dates.

[27]We convert ACT scores to SAT equivalents using SAT/ACT concordance data from Lavergne and Walker (2001), which relied on 1999/2000 data, and Dorans (1999), which relied on 1994-1996 data. Each source specifies a range of SAT scores corresponding to each ACT score. We construct an implied SAT equivalent for each source using their tables directly, and take the average of the implied scores across sources to generate our SAT/ACT correspondence. Results are not sensitive to the methodology. One additional school is dropped because the school-average ACT score in the SAT/ACT database is equal to 4.

offering the AMC. The average school has 0.8 students score at least 100 on the AMC 12 and 0.11 students score at least 120.[28] The number of female high scorers is substantially lower. Relative to the average public, non-charter, non-magnet, coed high school in the U.S., the average school offering the AMC is larger, has more Asian-American and fewer black and Hispanic students, is less likely to receive Title I funding, and has fewer students qualifying for the free lunch program. Sample AMC schools are also located in wealthier, more urban ZIP codes with more highly-educated adults. As expected given school and region demographics, AMC schools also perform better than the average U.S. public school on the SAT math, having higher mean scores and more students with perfect scores.

4 Differences in High Math Achievement Across Schools

In this Section, we bring out two basic facts. There are large systematic differences in the rates at which different schools produce high math achievers related to the schools' demographics. And there are also large differences among seemingly similar schools.

4.1 Achievement gaps

In this section we explore the magnitude of various achievement gaps among high-achieving math students by examining the relationship between the number of high-AMC scorers in a school and its demographics. The first column of Table 3 presents coefficient estimates (with standard errors in parentheses) from a negative binomial regression with the number of students in a school scoring at least 100 on the AMC 12 as the dependent variable. The estimates indicate that several observable characteristics of a school/neighborhood are strong predictors of the number of high math achievers that a school will produce. Parental education is very important: a one percentage point increase in the fraction of adults with bachelor's and graduate degrees increases the expected number of AMC high-scorers by 2.5 and 7.0 percent, respectively. Racial and ethnic composition also matters. The estimates suggest that a one percentage point increase in the Asian-American population increases the expected number of AMC high scorers by 1.8 percent. We also find that there are fewer high AMC scorers in schools that have more Hispanic students and those that have more low-income students qualifying for the free lunch program.

One demographic variable that would be significant in most regressions of school-mean

[28] In the full set of 3,105 schools that we matched to the NCES data the mean number of students scoring at least 100 on the AMC 12 is 0.93, reflecting that the dropped schools (which are mostly private, magnet, or magnet-like) have more high scorers per school.

standardized test scores on demographics that does not have the expected effect here is that the number of AMC high scorers is not higher in higher income areas. Several potential explanations for the lack of an income effect are possible; e.g., it could reflect nonlinearity in the income-AMC relationship (AMC-participating schools are disproportionately located in upper income areas) or it could be due to a selection effect with participating schools in low-income areas being highly nonrepresentative.[29]

4.2 Magnitudes of differences among seemingly similar schools

In this Section, we document that there are also large differences among seemingly similar schools. One way to do this is via the negative binomial regression estimates. One justification for the negative binomial model is if the number of high scorers in school i is Poisson with mean $e^{X_i\beta}u_i$, with u_i being a multiplicative gamma-distributed unobserved shock to a school's production rate that has mean 1 and variance α.[30] For example, a school would have a u_i of 0.5 if each of its students were only half as likely to score 100 on the AMC 12 as were students at the average school with comparable demographics, and a u_i of 1.5 if its students were 50% more likely to succeed than would be expected given the demographics. In our negative binomial regression of the number of students scoring at least 100 on the AMC 12 on school demographics, the estimated variance of the multiplicative random shock is $\hat{\alpha} = 0.73$. A variance of 0.73 corresponds to a standard deviation of 0.85, i.e. a school that is one standard deviation above average is producing AMC high-scorers at 185% of the average rate and a school that is one standard deviation below average is producing AMC high scorers at just 15% of the average rate. For the variance to be this large there must be a substantial number of schools producing AMC high scorers at a small fraction of the average rate and/or a number of schools producing AMC high scorers at two or more times the average rate.[31]

We conclude that demographic differences account for a substantial portion of the variation across schools in the number of students who achieve high AMC scores, but that there are also substantial differences across seemingly similar schools.

[29] The income effect becomes smaller in magnitude but remains negative and significant if we drop the Title 1 and free lunch variables.

[30] This model is sometimes referred to as the Negbin 2 model. See, e.g., Cameron and Trivedi (1986) and Greene (2008) for a discussion of negative binomial functional forms.

[31] The estimate is highly significant and a likelihood ratio test rejects the Poisson alternative at an extremely high significance level.

5 Distributions of "School Effects"

We noted above that the negative binomial model can be regarded as providing an estimate of the variance of the unobserved school effects that we would need in addition to the demographic differences to reproduce the observed heterogeneity in the counts of AMC high scorers. Excess variance, however, can take many forms: it may be due to a set of underachieving schools that produce very few high scorers, or to a small (or large) set of extreme (or not so extreme) overachieving schools, etc. In this Section, we provide estimates of the distribution of school effects that would lead to the distribution of outcomes observed in the data. One observation is that the distribution includes a thick tail of schools that produce many more AMC high scorers than one would expect given their demographics.

Suppose the number of AMC high scorers in school i, y_i, is distributed Poisson (λ_i), where $\lambda_i = e^{X_i\beta}u_i$, X_i is a vector of observable characteristics, and u_i is an unobserved "school effect" with a multiplicative effect on the Poisson rate. We assume that the school effect u_i has an unknown density f. Appendix A1 describes a methodology for estimating both the coefficients β on the observable characteristics and the distribution f of the unobserved school effects. In a nutshell, we estimate the density via a series estimator: we model $f(x)$ as a product of a gamma-like term, $x^\alpha e^{-x}$, similar to that used in the negative binomial model and an orthogonal polynomial expansion, note that different coefficients on the polynomial terms produce different likelihoods of seeing 0, 1, 2, etc. high scorers, and use maximum likelihood estimation to find a density that comes closest to matching the observed frequencies of the outcomes conditional on the demographics.

We estimate the model on the same dataset as the negative binomial regression of the previous Section: we use the count of students scoring at least 100 on the AMC 12 as the dependent variable and include the same set of demographic controls. The second column of Table 3 reports the estimated coefficients on the demographic controls. They are similar to the negative binomial estimates in the first column, indicating that those estimates are robust to the more flexible modeling of the unobserved heterogeneity. Indeed, a comparison of the maximized log-likelihoods at the bottom of the Table shows that the semi-parametric model fits the data only slightly better than the negative binomial model discussed in the previous Section, suggesting that the true distribution of the u_i is fairly similar to a gamma distribution.[32]

[32]See Table 5 in the Online Appendix for a comparison of predicted counts under the Poisson, negative binomial, and semi-parametric models to the actual counts in the data. The Table includes, for each model, a χ^2 test of goodness-of-fit; the tests confirm our visual assessment that the actual data reject the Poisson model at a high level of significance, but they do not reject the negative binomial or semi-parametric models

Our primary interest here is in the distribution of the school effects. We note first in Table 3 that relaxing the assumption that the u_i are gamma-distributed increases the estimated variance from 0.73 to 0.96. The top panel of Figure 1 graphs the probability density function from which the unobserved school effects u_i are estimated to be drawn. The x-axis corresponds to different possible values of the unobserved effect; e.g., a value of $u = 1$ corresponds to a school that produces AMC 12 high scorers at exactly the mean rate given its demographics, a value of $u = 0.5$ corresponds to a school that produces high-scorers at half of this rate, etc. Informally, the curve is like a histogram giving the relative frequency of the values of u in the population of schools. Substantial differences between schools with similar demographics are evident: the distribution is not tightly concentrated around $u = 1$. Instead, there are a large number of schools that produce AMC 12 high scorers at well below the average rate; e.g., about 32% are estimated to produce high scorers at less than half of the average rate. At the other end, there are many highly successful schools producing high scoring students at 50 to 100% above the average rate. The dashed lines in the figure are 95% confidence bands for the estimated density.[33] They indicate that the estimates are fairly precise throughout most of the range.

A striking feature of the distribution that is not immediately apparent from the PDF graph is that the estimates indicate that there is a thick upper tail of extremely successful schools. The bottom panel of Figure 1 illustrates this better by graphing in bold the CDF of the estimated distribution for u's ranging from 2 to 8. The estimates indicate that 11% of schools are estimated to be producing high scorers at more than twice the average rate and there is a substantial mass (about 1%) producing high achieving students at more than five times the average rate for a school with their demographics. The dashed lines again give a 95% confidence interval. They indicate that the thick tail is a statistically significant phenomenon.

6 Challenges to the Interpretation of the School Effects

As we noted earlier, the "school effects" we have estimated conflate multiple factors. They will reflect differences in causal effects of school environments on potential high math achievers. But they will also reflect other less interesting sources of outcome heterogeneity: demographic differences not captured by variables in our dataset; differences in unobserved

at conventional levels, having p-values of 0.7 in each case.

[33] The confidence bands in this figure were generated using the parametric bootstrap procedure described in the Online Appendix. We also generated confidence bands using the nonparametric bootstrap procedure described there. They are quite similar (though slightly wider).

student ability attributable to location decisions made by parents of gifted children; and even less interestingly, differences in the fraction of the high-achieving math students at each school who take the AMC 12 test. In this Section, we present several auxiliary estimates aimed at providing some insights on the importance of these less interesting sources of outcome heterogeneity. We will argue that they do not seem sufficient to account for the variation we have found.

6.1 Selection into test-taking: evidence from extreme high achievers

A portion of the "school effects" we have reported will be due to differences in the AMC 12 participation rates for high-achieving math students from different schools. The primary way in which we can provide some evidence of whether this could be driving our results is to provide additional estimates derived from counts of students at even higher achievement levels.

Students scoring at least 120 on the AMC 12 can be thought of as well above the 99.9^{th} percentile among college bound students. The unique ability of the AMC 12 to distinguish among such extreme high achievers makes it possible to examine their agglomeration as well. We think that there are many students at AMC-offering schools who would have scored 100 if they had taken the AMC 12, but who did not take the test. We believe, however, that the fraction of students at AMC-offering schools who would have scored 120 if they took the AMC 12, yet chose not to participate, is much smaller. Hence, we can compare results obtained with an AMC 12 cutoff of 100 to results obtained with an AMC 12 cutoff of 100 to see if results change when selection into test-taking becomes less important. Moreover, we believe that the issue of selection into test-taking is small in absolute terms at the 120 level and hence results with the 120 threshold cannot be greatly affected by selection into test taking. We believe this for a few reasons. First, scoring 120 on the AMC 12 requires both a great deal of natural ability and a lot of effort dedicated to learning high school mathematics very well and we feel that it is unlikely that students would have made the effort if they were not interested in participating in math competitions. 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.[34] Second, we can provide some statistical evidence from looking at repeat test-takers across

[34] Anecdotally, we have discussed discoveries of star students with many math team coaches. Many have stories that involve students they had not known showing up to an AMC or some other test and doing very well. None, however, involved an initial encounter in which a student did something as impressive as scoring 120 on the AMC 12.

years. Considering the set of students who were among the top 1% of 11th graders on the 2006 AMC 12 and attended a school that participated in the 2007 AMC 12, we are able to identify 80% as taking the AMC 12 in 2007.[35] Third, we can look at students who received other math honors and see if they had taken the AMC 12. In the 2007 Intel Science Talent Search five students were named as finalists on the basis of having done outstanding mathematical research projects. We know from the published lists of AMC winners that all five took the 2007 AMC 12. Of the 30 winners or honorable mentions on the Georgia math contest mentioned earlier, we know that 100% (all 30 of 30) took the 2007 AMC 12. In the case of the Massachusetts contest mentioned earlier, 17 of the 20 winners took the 2007 AMC 12.[36] These comparisons suggest that the number of non-takers in our schools who would have scored 120 is at most 10% to 20% of the number who did score 120.

The third column of Table 3 presents estimates from a negative binomial regression using school-level counts of students scoring at least 120 on the AMC 12 as the dependent variable. The coefficients on parental education, income, and racial/ethnic variables are all quite similar to those derived from counts of students scoring at least 100 on the AMC 12, though the point estimates are generally larger in magnitude. None of the differences are statistically significant, with the exception of the coefficient on the fraction of adults with bachelor's degrees. The most important estimate for our current purposes is that for the parameter $\hat{\alpha}$, the estimated variance of the unobserved school effects u_i. The estimate of 2.18 not only remains highly significant in this environment in which we think selection into test-taking is unimportant, but is substantially larger than the estimate from the regression run at the AMC 100 level. This bolsters the case that the earlier results were not primarily driven by differences in participation rates. And it provides a new striking result on extreme high math achievement: school environments appear to be even more important in influencing whether students will reach this very high level.

6.2 Unobserved demographic differences: evidence from SAT/ACT high scorers

Another portion of the "school effects" we have reported will be due to unobserved demographic differences. For example, a school may do well because many of its parents with

[35] This statistic underestimates participation because we have no way to match students who wrote their names differently or changed schools from one year to the next. To get some sense of what might be done with manual matching and local knowledge, we manually matched all students from Massachusetts who scored at least 110 on the 2013 AMC 12A and were still in high school to the 2014 published AMC 12 winners lists. Here, we found 11 of the 12 2013 high scorers on the 2014 winners list.

[36] The other three winners also participated in the AMC series, but were younger and chose to take the AMC 10 rather than the AMC 12.

graduate degrees are Ph.D.s in mathematical and technical fields, or because it attracts many parents of high-ability children (perhaps because its district has a gifted program that is highly regarded even if it is not effective). In this Section, we present some evidence on the magnitude of unobserved demographic differences by estimating school effects using counts of students achieving perfect scores on the SAT and ACT math tests.[37]

The fourth column of Table 3 presents estimates from a negative binomial regression with the same demographic controls as before. The most important estimate for our current purposes is again the parameter $\hat{\alpha}$ giving the estimated variance of the unobserved school effects u_i. The estimate of 0.23 is statistically significant at the 0.1% level, indicating that there are unobserved demographic differences and/or differences in how well the schools in our sample prepare their students to get very high math SAT/ACT scores. But the magnitude of the coefficient here is much smaller than the estimates of 0.73 and 2.18 we had obtained when looking at students scoring 100 or 120 on the AMC 12. This suggests that the differences in the counts of AMC high scorers are not primarily due to unobserved differences in demographics or student abilities. One story that would be consistent with both results is that there might be more heterogeneity in the extent to which schools encourage students to develop the deeper understanding of high school mathematics needed to perform well on the AMC: most schools see it as their responsibility to teach students the math that appears on the SAT but there may be more heterogeneity in whether schools feel that it is important to offer additional enrichment to gifted math students.

Figure 2 presents estimated distributions of school effects from the data on SAT/ACT high scorers. The estimated PDF in the top part of the figure has one clear difference from the PDF of the AMC school effects: the distribution is much closer to being symmetric about the $u = 1$ mean whereas the AMC distribution was skewed to the right. One implication is that there are significantly fewer (14% vs. 32%) schools that are more than 50% below average in production of high scores on the SAT/ACT than there are at producing high scores on the AMC. A second striking difference is apparent in the bottom panel: the SAT/ACT distribution has a much thinner upper tail. In the SAT data, only 2% of schools are estimated to produce high scorers at more than twice the average rate, and just 0.02% are estimated to produce high scorers at more than five times the average rate. This contrasts with our earlier estimates that 11% of schools that were estimated to produce AMC 12 high scorers at more than twice the average rate and 1.0% at more than five

[37]Recall that we prioritize SAT scores in this calculation, counting a student who took both exams as having a perfect score if and only if he or she had a perfect score on the math portion of the SAT reasoning test.

times the average rate. We interpret this contrast as suggesting that the thick upper tail in the AMC 12 school effects distribution is not primarily due to differences in unmeasured student characteristics.

The estimated coefficients on the demographic variables are generally quite similar in the SAT/ACT and AMC estimations, and the former are again similar regardless of whether the coefficients are obtained from negative binomial regression, shown in the fourth column of Table 3, or from our semi-parametric estimation, shown in the fifth column. The fact that observed demographics affect the AMC and SAT regressions similarly suggests that unobserved demographic differences may also have similar effects in the two regressions. Assuming this to be the case, the fact that the estimated $\hat{\alpha}$ in the SAT/ACT regression is so much smaller than that in the AMC regression would imply that at most 32% of the variance in the AMC school effects is due to unobserved demographic differences. We regard this as a conservative bound because the SAT/ACT school effects reflect more than demographic differences – we assume that there are idiosyncratic differences in how well schools prepare students for the SAT/ACT – and these will be part of what is captured by the SAT/ACT school effects.

6.3 Unobserved demographic differences: evidence from gender differences

The effects of schools on female students with high math ability are of independent interest given the underrepresentation of women in mathematical and technical fields. Data on the female high-scorers also provides another potential source of information into whether heterogeneous outcomes are driven by unobserved demographic differences: differences such as whether a district has many parents with Ph.D.s should be similarly relevant to male and female students (provided there are not large differences in how parents of high ability girls and boys choose where to live). A number of plausible explanations could be given for why there might be more agglomeration of high-achieving girls. For example, the dispersion of school effects would be larger for girls if there is variation in how encouraging/discouraging schools are toward girls independent of a general school-quality effect. Or peer effects could be more important for girls. Or the rigor of a school's classes might be more important for girls because they are less liable to complain or take supplementary online classes.

The last column of Table 3 reports coefficient estimates from a negative binomial regression with a count of the number of female students scoring at least 100 on the AMC 12 as the dependent variable. Note that the estimated variance of the unobserved school

effects, $\hat{\alpha}$ of 0.95 (s.e. 0.29) whereas it was 0.73 (s.e. 0.08) when we examined high-scoring students of either gender. This indicates that there may be more underlying variation in the rate at which different schools are producing high-achieving girls. The substantial noise in the variance estimates from the girls-only sample, however, is such that we cannot say whether the difference in variances is significant.

7 Mechanisms Behind the School Effects

In the preceding sections we argued that a substantial portion of the idiosyncratic differences in the rates at which seemingly similar schools produce high-achieving math students are due to some sort of environmental differences. In this section we present several additional analyses designed to provide insight into what may be leading to these differences.

7.1 Effects on high-achieving students or generally strong math programs?

The fact that a school produces many high-achieving students need not imply that the school's environment particularly benefits high-achieving students: it could be that the school just has a generally strong math program. If, for example, the math program at school i raises the score of each student j from θ_j to $\theta_j + \Delta_i$, then a school with a larger Δ_i will have more students scoring above any particular threshold.

To explore whether this appears to be a large part of what is going with our AMC results, we obtained data on the average math SAT/ACT score for students within each school, which we think of as reflecting the general quality of the school's math program (as well as observed and unobserved demographics). We then repeated our negative binomial regressions of the number of students scoring at least 100 and at least 120 on the AMC 12 on the same demographics as before plus two additional variables: the average SAT/ACT score in the school and the SAT/ACT participation rate. We find that the added variables only moderately reduce the estimated variance of the school effects. In the AMC ≥ 100 regression the estimated variance $\hat{\alpha}$ drops from 0.73 (s.e. 0.08) to 0.57 (s.e. 0.07). In the AMC ≥ 120 regression the estimated variance $\hat{\alpha}$ drops from 2.18 (s.e. 0.50) to 1.60 (s.e. 0.41).

We also perform a similar exercise in our regressions examining counts of students with perfect SAT/ACT scores.[38] In the SAT/ACT scores, including this measure of the general

[38]The SAT/ACT participation rate is already included in the controls, so in this exercise we simply add mean SAT/ACT score as a covariate.

quality of the math program reduces the unobserved heterogeneity nearly to zero – the estimated variance $\hat{\alpha}$ drops from 0.23 (s.e. 0.03) to 0.07 (s.e. 0.02).

If one thinks of the difference between the $\hat{\alpha}$ estimated from the AMC data and the $\hat{\alpha}$ estimated from the SAT/ACT data as a conservative estimate of the variance in in the school effects that controls for both observed and unobserved demographic differences, then the finding of this section is that this difference is 0.50 ($= 0.73 - 0.23$) when one does not control for the school-average SAT score and also 0.50 ($= 0.57 - 0.07$) when one does.[39] We conclude that a substantial portion of the "school effects" we have reported seems to be due to factors that differentially impact high-achieving students.

7.2 Peer effects or differences in school quality?

Although we have sometimes described our estimated "school effects" as reflecting the heterogeneous rates at which schools produce high scorers, it is more accurate to describe them as a quantification of the excess agglomeration of high-achieving students. Agglomeration will occur if there is unobserved heterogeneity in school "quality." But it will also be present if there are peer effects among high-achieving students. In this Section, we provide a formal nonidentification result, noting that one cannot distinguish peer effects from school quality differences using data on a single cross section; we then show that a calculation using data from multiple years suggests that a portion of the school effects we have found are due to peer effects, but that a larger portion is not.

The impossibility of distinguishing peer effects from school quality differences can be formalized using standard results on the binomial distribution. First, consider a model with no peer effects in which schools differ in unobserved quality (which is captured by a gamma-distributed random variable):

Model 1 *Suppose the count of high scorers* $Y_i \sim Poisson(\lambda_i)$ *with* $\lambda_i = e^{X_i\beta}u_i$ *where* $u_i \sim \Gamma(\frac{1}{\alpha}, \frac{1}{\alpha})$.[40]

Second, consider a model with no unobserved heterogeneity u_i in school quality, but with peer effects between high-achieving students. Specifically, consider a model in which high scorers are produced in two ways: the school directly produces high-scoring students at

[39] This comparison is essentially unchanged when we include richer controls for school-average SAT score and participation. The estimated variance $\hat{\alpha}$ in the AMC ≥ 100 regression drops from 0.57 (s.e. 0.07) when only linear controls are used to 0.56 (s.e. 0.07) when cubic polynomials of mean SAT/ACT and participation, plus an interaction between mean SAT/ACT and participation, are included; the equivalent change in the SAT/ACT high-scorers regression is from 0.07 (s.e. 0.02) to 0.04 (s.e. 0.01).

[40] The density of the assumed distribution of the u_i is $f(u) = (1/\alpha)^{\frac{1}{\alpha}} e^{-\frac{1}{\alpha}u} u^{\frac{1}{\alpha}-1}/\Gamma(1/\alpha)$.

a Poisson rate; and high scorers produce additional high scorers via an infection-style dynamic.

Model 2 *Suppose a school directly produces high scoring students at Poisson rate* $\lambda(X_i)$ *during the time interval* $[0, 1]$. *Suppose that in each subinterval* $(t, t+dt)$, *each high scoring student then present produces another high-scoring student with probability* $g(X_i)dt$. *Let* Y_i *be the number of high scoring students at* $t = 1$.

The two models are well-known to produce counts that follow the negative binomial distribution.[41] As a result, we cannot distinguish between the two models given a dataset containing a single observation on each school. Conceptually, the argument is similar to Ellison and Glaeser's (1997) argument that unobserved comparative advantages and spillovers can lead to equivalent geographic concentration.

Proposition 1 *The distribution of* $Y_i|X_i$ *under Model 1 with parameters* (α, β) *is identical to the distribution of* $Y_i|X_i$ *under Model 2 if the direct production rate is* $\lambda(X_i) = \frac{1}{\alpha}\log(1 + \alpha e^{X_i\beta})$ *and the peer infection rate is* $g(X_i) = \alpha\lambda(X_i)$.

While the peer effects formulas may seem complicated at first, one can think of them as saying that it is the ratio of the peer infection rate $g(x)$ to the direct production rate $\lambda(x)$ that determines the magnitude α of the excess variance (relative to what one would expect with only direct production). Using the approximation $\log(1 + y) \approx y$, one can think of the formula for the direct production rate as $\lambda(X_i) \approx e^{X_i\beta}$, which is the same functional form as in the unobserved heterogeneity model with the unobserved component set equal to its mean.

A model in which there are both school quality differences and peer effects of the above form will not produce an exact negative binomial distribution, but the excess variance will still be related to the amount of heterogeneity and the strength of the peer effects in a similar manner. Formally, consider a hybrid model in which school i directly produces high scorers at Poisson rate $\lambda_i = \frac{1}{\alpha_p}\log(1 + \alpha_p e^{X_i\beta}u_i) \approx e^{X_i\beta}u_i$ during the time interval [0, 1], with u_i being a gamma-distributed random variable with mean 1 and variance α_u. As in Model 2, suppose that each high scorer produces additional high scorers at Poisson rate $g_i = \alpha_p\lambda_i$ and let Y_i be the number of high scorers at $t = 1$. A calculation gives

Proposition 2 *In the hybrid model we have* $E(Y_i|X_i) = e^{X_i\beta}$ *and* $Var(Y_i|X_i) = E(Y_i|X_i) + \alpha E(Y_i|X_i)^2$ *for* $\alpha = \alpha_u + \alpha_p + \alpha_u\alpha_p$.

[41] In Model 1, $Y_i \sim NB\left(\frac{1}{\alpha}, \frac{\alpha e^{X_i\beta}}{1+\alpha e^{X_i\beta}}\right)$. In Model 2, $Y_i \sim NB\left(\frac{\lambda(X_i)}{g(X_i)}, 1 - e^{-g(X_i)}\right)$. See Boswell and Patil (1970), Section 8.2, or Karlin (1966), p. 345 for proofs.

Hence, although a negative binomial model will be misspecified, one way interpret an excess variance parameter α estimated from count data is as a reflection of $\alpha_u + \alpha_p + \alpha_u\alpha_p$, which consists of a sum of the strengths of the two agglomerative forces plus an interaction term.

Suppose now that we are able to observe two conditionally independent draws Y_{i1}, Y_{i2} for each school. By "conditionally independent" we mean that the school characteristics X_i and unobserved quality u_i are the same at both $t = 1$ and $t = 2$, but that the subsequent Poisson realizations are independent and that peer infections operate separately within each time period. Define $\overline{Y}_i = Y_{i1} + Y_{i2}$ to be the sum of the counts of high-scoring students across the two draws. Suppose that with such data one estimates two excess variance parameters: first, treat the Y_{it} as $2N$ observations and estimate an excess variance parameter α; and second, treat the $\overline{Y}_i$ as N observations and estimate an excess variance parameter $\overline{\alpha}$. A result relating the estimates to the relative importance of peer effects and unobserved heterogeneity is

Proposition 3 *Suppose α and $\overline{\alpha}$ satisfy $0 < \frac{\alpha}{2} < \overline{\alpha} < \alpha$. Then there is an unique pair of parameters for the hybrid model (α_u, α_p) for which $Var(Y_i|X_i) = E(Y_i|X_i) + \alpha E(Y_i|X_i)^2$ and $Var(\overline{Y}_i|X_i) = E(\overline{Y}_i|X_1) + \overline{\alpha}E(\overline{Y}_i|X_i)^2$. Specifically, this holds for $\alpha_u = 2\overline{\alpha} - \alpha$ and $\alpha_p = \frac{2(\alpha-\overline{\alpha})}{1+2\overline{\alpha}-\alpha}$.*

The above result implies that the reduction in overdispersion that results when we sum two observations per school will let us infer the relative importance of peer effects and unobserved heterogeneity in generating the overdispersion. Intuitively, if the excess variance is due to unobserved heterogeneity in school quality that does not change over time, then the combined data from two years should show just as much overdispersion as a single year of data. But if the overdispersion is due to within-time-period peer effects, then the overdispersion will decline as we combine results from multiple years. It should be kept in mind, of course, that the hybrid model uses extreme assumptions: the unobserved school effects u_i are assumed to be perfectly persistent; and peer effects are not felt across time periods. In practice, school effects would be expected to be imperfectly correlated across time as teachers leave, curricula change, etc.; and peer effects may be relevant even between students who are never in school together via chains where student A infects student B who later infects student C, etc. An application of Proposition 3 will overestimate the importance of peer effects if the former factor is more important and underestimate it if the latter dominates.

To investigate the relative importance of unobserved school effects and peer effects in the AMC data we combine the dataset we have examined so far with a comparable dataset

containing counts of the number of students in each school scoring an equivalent of 100 on the 2003 AMC 12.[42] The four-year interval between observations should make observations roughly conditionally independent in that the sets of students in the high school in the different test years are nearly disjoint. It should also eliminate many cross-period peer effects although it is possible that a student who achieved a high score in 2003 influenced a student still in high school in 2007 either directly if the students overlapped at the school at some point in 2004-2006 or indirectly via some chain of influence. We hope that the four-year interval is also short enough so that unobserved school qualities will be similar across the two years. We restrict our attention to the set of public, nonmagnet, noncharter, public schools which offered the AMC 12 in both 2003 and 2007. The subsample includes 1,606 of the 1,984 schools in our previous analyses.

We perform two negative binomial regressions on the combined dataset. First, we run the regression with each school's 2003 and 2007 high scorer counts being treated as two independent observations. The estimated α in this model is 0.74, which is similar to that we found earlier in the 2007 data. Second, we estimated a negative binomial regression with just one observation per school using the combined count $\overline{Y}_i \equiv Y_{i,2003} + Y_{i,2007}$ as the dependent variable. The estimated $\overline{\alpha}$ in this model is 0.65. Using the formula in the proposition above we find that the parameters mutually consistent with the two estimates are $\alpha_u = 0.56$ and $\alpha_p = 0.12$. We conclude that some relatively permanent factor appears to be more important than within-cohort peer effects (or transitory school quality) in producing the observed clustering across schools. Again, however, we should emphasize that part of the permanent factor could be due to some sort of peer effect of a different sort than is normally considered; e.g., it could be due to a community spirit that develops within a school and is passed down from one cohort to the next.

Estimates from a regression examining higher-achieving students scoring an equivalent of 120 on the AMC 12 and from analyses of high-achieving females only are similar. For the higher achievement threshold of 120 (and its equivalent of 126.5 in 2003), the estimated α and $\overline{\alpha}$ are 1.93 and 1.49. The parameters mutually consistent with the two estimates are $\alpha_u = 1.05$ and $\alpha_p = 0.43$. For female students scoring 100 or higher (111.5 or higher in 2003), the estimated α and $\overline{\alpha}$ are 0.97 and 0.84. The parameters mutually consistent with the two estimates are $\alpha_u = 0.71$ and $\alpha_p = 0.15$. Again, in each case, the persistent factor

[42] As before differences in difficulty across tests make it desirable to adjust the cutoff when using different tests. We use a cutoff of 111.5 rather than 100 on the 2003 AMC 12A because that makes the fraction of students scoring at least equal to the cutoff as close as possible to the fraction scoring 100 on the 2007 AMC 12A.

affecting performance across years appears to dominate the inferred strength of transitory peer effects.

7.3 Informal evidence on high-achieving schools

To get additional insight into what upper-tail schools might be doing to promote high math achievement, we present some informal descriptive evidence about schools that produce an unexpectedly large number of AMC high scorers.[43] Specifically, Table 4 presents data on 20 high-achieving schools along with sample means for these schools and for 20 comparison schools.[44] The high-achieving schools averaged 6.9 students scoring at least 100 on the AMC 12, whereas the comparison schools averaged 1.0.

One initial comment about the high-performing schools is that most are ordinarily situated public high schools. One, Oak Ridge HS in Oak Ridge, TN, is located near a national laboratory. A second, Cardozo HS, is located within New York City, which has extensive school choice. But most do not seem unusual and more than half are either the unique comprehensive high school in their school district, or one of just two or three comprehensive schools in districts that primarily divide students geographically and offer similar programs at each of their schools.

Comparing the summary statistics we note several differences between the high-achieving schools and the matched comparison group. One clear difference is that the high-achieving schools were much more likely to have "star" math teachers.[45] In some cases star teachers seem extremely important (and impressive). For example, for over 40 years Lincoln East's Leona Penner taught their top math students for four years in a row from 7th through 10th grade and followed a special curriculum that focused "more on number theory, problem solving, logic and proof than the traditional curriculum."[46] Popular press stories suggest

[43] We selected 20 schools for which $E(u_i|y_i)$ is largest when we assume that the school effects u_i are independent draws from the distribution estimated under our semiparametric model, and y_i, the count of students scoring at least 100 on the AMC 12, is also assumed to be generated as in the model with the estimated parameters. Note that in order to have a high posterior mean, observations from the school will need to be highly informative, which results in these schools tending both to have a high ratio of actual to predicted high scorers and a large number of high-scorers.

[44] For each school, we chose as a comparison the school in the same state which was most similar demographically in the sense of minimizing $|X_i - X_j|'|\hat{\beta}|$.

[45] To define this variable we labeled a school as having a star math teacher if a math teacher's name appeared on the school's Wikipedia page or if Google searches revealed that the math team coach was active beyond the school environment, was the subject of a glowing portrayal on a nonaffiiated site, or had won highly prestigious honors such as being on the USA Today All-USA Teaching Team.

[46] Her middle school math teams won the Nebraska state Mathcounts championship in 26 of her last 29 years and she won a number of other awards (Reist, 2012). Other long-serving stars include Canton's Martin Badoian, whose math teams won 19 state championships in one 21 year period and who still teaches at age 86 (Redd, 2004), and Vestavia Hills' Kay Tipton, who founded their math team in 1975 and went on to win

a combination of reasons for other stars' success including knowledge, work ethic, and a remarkable facility for motivating students to put in extra effort. For example, in discussing how Honey Creek's Robert Fischer had managed to coach teams to 17 state and one national championship in math, 21 state and 4 national championships in chess, and 35 county championships in tennis, USA Today noted that he "starts his day at 5:30 am and draws 35-50 students for an hour of before-school math; holds Lunch Math through lunch periods; coaches Mathcounts and tennis after school" (Schneider, 2001).[47]

There also appear to be differences in the curricular offerings. In looking at the most standard high school courses, geometry and algebra II, we found that the high-achieving schools tend to stratify their math offerings more finely: they typically offer three levels of these classes whereas the comparison schools usually offer two. They are also more likely to offer multivariable calculus: we found such courses at five of the high achieving schools and just two of the comparison schools. The final column records observations of additional special curricular offerings. The "-1" notation marks five schools which offer some type of additional math course which includes problem solving and/or math competition in its description. Often the descriptions indicate that the classes offer a variety of types of enrichment.[48] The most striking example is that of the top-ranked school, Vestavia Hills HS, at which students may in addition to their regular math course enroll each year in an extra one-half or full-credit "Honors Math Theory" class that meets every day before school and/or during lunch. The "-2" notation marks schools that offer classes to prepare students for research competitions, which is something we found at two high-achieving and one comparison school.

Again, we should note that in dropping schools with magnet-like programs we are omitting many high-achieving schools where it is easy to identify institutions designed to promote high achievement. For example, the Rockdale Magnet School for Science and Technology, which we mentioned earlier, has a number of unusual curricular offerings, including a two-year sequence covering single and multivariable calculus with linear algebra and differential equations, a history of mathematics elective, and (in some years) an extra math team course. The courses were designed by a star teacher, Dr. Charles Garner, and

the Mu Alpha Theta national championship at least 14 times ("Math Teacher Kay Tipton", 2013).

[47] Fischer teaches at Honey Creek Middle School, which feeds into Terre Haute South Vigo.

[48] For example, Georgetown's Independent Study in Mathematics course description says, "This course will extend mathematical understanding beyond the Algebra II level in a specific area or areas of mathematics, such as theory of equations, number theory, non-Euclidian geometry, discrete mathematics, advanced survey of mathematics, or history of mathematics. This course will provide students opportunities to pursue interest in mathematical topics via independent research, directed learning, preparation for and participation in challenging mathematics competitions, and/or mentoring by a mathematics professional."

the school also has an active math team.[49] In a personal communication, Dr. Garner attributed his students' success on the AMC to "the hard work they do to develop these problem-solving skills" and noted that "very, very few of our students walk into 9th grade with a love of math!"

In summary, a quick look at the high-achieving schools suggests that multiple factors may be involved in these schools' success. In several cases a star teacher may be playing an important role. And there also seem to be institutional differences including decisions by the schools to stratify classes more finely and offer additional classes that provide enrichment and/or contest preparation in addition to the standard high school course sequences.

8 Conclusion

In this paper we have used data on the Mathematical Association of America's AMC 12 exam to provide a look at high-achieving math students in U.S. high schools. Our most basic observation is that they are highly agglomerated. Much of this is associated with strong demographic predictors of high math achievement: even though our sample consists mostly of relatively high-performing schools in relatively affluent areas we note that both the presence of Asian-American students and parents with advanced degrees are strong predictors of which schools will produce high-achieving students. But beyond this we find that there are also large differences among seemingly similar schools.

As a first step in exploring these differences we estimated the distribution of the unobserved "school effects" that would be needed to produce the observed patterns. Methodologically, we note that this distribution can be estimated (and estimated fairly precisely in our data) even though almost all schools only have a handful of high-scorers. The most interesting aspect of the estimated distribution is that it has a thick upper tail. Many (about 200) of the schools our sample appear to be producing AMC high scorers at more than twice the expected rate and some (about 20) are producing them at 5 to 10 times the expected rate.

We have presented a number of pieces of auxiliary evidence to suggest that there are real differences in school environments. To examine selection into test-taking we take advantage of the AMC's ability to identify students at even higher percentiles where we think selection is not important and note that school effects appear to be even more important. To get some

[49] Dr. Garner has been active outside his school both with the Georgia Department of Education and in math competition communities, has won various awards, and written a Calculus textbook and edited several other books.

idea of the portion of our school effects which are due to unobserved differences in student ability we estimated the dispersion in the rates at which the schools in our sample produce students with high SAT/ACT scores. We found that the variance was only 30% as large. To the extent that some portion of the SAT school effects are due to differences in school quality, the message that we are not primarily finding unobserved demographic differences becomes even stronger. Also, no comparable upper tail is visible in the SAT/ACT data.

Our primary focus is on documenting that there appear to be substantial differences in the rates at which different schools produce high math achievers. In our final section we tried to provide some additional evidence into mechanisms that may be involved. Methodologically, we bring out how peer effects and differences in school quality can produce agglomeration patterns that are indistinguishable given data in a single cross-section but potentially distinguishable given data on both within cohort agglomeration and cross-cohort coagglomeration. Our estimates here suggest that there are some peer effects especially at the highest level of students scoring 120 on the AMC 12 (or that a portion of the school effects are time varying). This may be an interesting result on its own. But, the more important message is that majority of the effect is something that is a more permanent feature of schools (although it could also be some other peer effect such as a sense of community that reproduces itself from one cohort to the next).

Our finding that schools matter for high achievement can be seen as contrary to the findings of several recent studies that particular gifted programs appear to have little impact on marginal students. Our view, however, is that there need not be any conflict here. The previous literature mostly focuses on how elite programs affect marginal students whereas we are focused on how they affect students at the extreme opposite tail. Our evaluation metric is also different in that we are focused on in depth understanding and advanced problem solving skills. Something we have not emphasized much is that there is a potential complementarity to bridging the gap between their school samples and ours – in dropping private schools and magnet programs from our sample we have dropped the schools attended by over 40% of the AMC high-scorers from our sample. Accordingly, getting more of a sense of whether magnet programs are effective on their most able students would be of great practical importance.

Our findings about the thick upper tail are intriguing because they suggest that the US could dramatically increase the number of high-achieving math students it produces. Of course, this would require that the effects be due to environmental differences and the environments are something that could be reproduced. But given the importance of high-

achieving students to the economy and the potential benefits to students we hope that our paper will spur future research on these topics. The potential gains may also be much larger than what one sees in the schools we have studied. Ninety percent of US high schools do not offer the AMC. If we were able to examine the rates at which students in those schools develop high-level math skills we might have found much larger differences and many students who might have reached high achievement levels if they had attended a typical school in our sample.

References

Abdulkadiroglu, Atila, Joshua Angrist, Susan M. Dynarski, Thomas J. Kane, and Parag Pathak (2011): "Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots," *Quarterly Journal of Economics* 126 (2), 699-748.

Abdulkadiroglu, Atila, Joshua Angrist, and Parag Pathak (2014): "The Elite Illusion: Achievement Effects at Boston and New York Exam Schools," *Econometrica*, 82(1), 137-196.

Altonji, Joseph (1995): "The Effect of High School Curriculum on Education and Labor Market Outcomes," *Journal of Human Resources* 30, 409-438.

Altonji, Joseph and Richard Mansfield (2010): "The Contribution of Family, School and Community Characteristics to Inequality in Education and Labor Market Outcomes," mimeo, Yale University.

Altonji, Joseph, Erica Blom, and Costas Meghir (2012): "Heterogeneity in Human Capital Investments: High School Curriculum, College Major, and Careers," *Annual Review of Economics* 4, 185-223.

Andreescu, Titu, Joseph A. Gallian, Jonathan M. Kane, and Janet E. Mertz (2008): "Cross-Cultural Analysis of Students with Exceptional Talent in Mathematical Problem Solving," *Notices of the American Mathematical Society*, 55 (10), 1248–1260.

Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer (2002): "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment," *American Economic Review* 92(5), 1535-1558.

Angrist, Joshua and Miikka Rokkanen (forthcoming): "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away from the Cutoff," *Journal of the American Statistical Association.*

Avery, Christopher, Caroline Hoxby, Clement Jackson, Kaitlin Burek, Glenn Poppe, and Mridula Raman (2006): "Cost Should Be No Barrier: An Evaluation of the First Year of Harvard's Financial Aid Initiative," National Bureau of Economic Research Working Paper 12029.

Bayer, Patrick and Christopher Timmins (2007): "Estimating Equilibrium Models of Sorting Across Locations, *Economic Journal* 117 (518), 353-374.

Boswell, M. T. and G. P. Patil (1970): "Chance Mechanisms Generating the Negative Binomial Distributions," in G. P. Patil (ed.) *Random Counts in Scientific Work. Volume 1.* State College: Penn State University Press.

Bowen, William, Martin Kurzweil, and Eugene Tobin (2005): *Equity and Excellence in American Higher Education.* Charlottesville: University of Virginia Press.

Bui, Sa, Steven G. Craig, and Scott Imberman (2014): "Is Gifted Education a Bright Idea: Assessing the Impact of Gifted and Talented Programs on Achievement," *American*

Economic Journal: Economic Policy, 6(3): 30-62.

Cameron, A. Colin and Pravin K. Trivedi (1986): "Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests," *Journal of Applied Econometrics* 1, 29-54.

Cameron, A. Colin and Pravin K. Trivedi (1998): *Regression Analysis of Count Data.* Cambridge: Cambridge University Press.

Card, David and Alan B. Krueger (1992): "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy* 100 (1), 1-40.

Coleman, James S. (1966): "Equality of Educational Opportunity Study (EEOS)." ICPSR0 6389-v3. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Dee, Thomas, and Brian Jacob (2011): "The Impact of No Child Left Behind on Student Achievement," *Journal of Policy Analysis and Management* 30(3), 418-446.

Dobbie, Will and Roland G. Fryer, Jr. (2014): "The Impact of Attending a School with High-Achieving Peers: Evidence from the New York City Exam Schools," *American Economic Journal: Applied Economics*, 6(3), 58-75.

Dobbie, Will and Roland G. Fryer, Jr. (2013): "Are High Quality Schools Enough to Close the Achievement Gap? Evidence from a Social Experiment in Harlem," *American Economics Journal: Applied Economics*, forthcoming.

Dorans, Neil J. (1999). *Correspondences Between ACT and SAT I Scores*, College Board Report No. 99-1, New York, 2.

Duranton, Gilles and Henry G. Overman (2005): "Testing for Localisation Using Micro-Geographic Data," *Review of Economic Studies* 72 (4), 1077-1106.

Efron, Bradley and Robert J. Tibshirani (1993): *An Introduction to the Bootstrap.* New York: Chapman and Hall.

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

Ellison, Glenn, Edward L. Glaeser, and William Kerr (2010): "What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns," *American Economic Review* 100 (3), 1195-1213.

Ellison, Glenn and Ashley Swanson (2010): "The Gender Gap in Secondary School Mathematics at High Achievement Levels: Evidence from the American Mathematics Competitions," *The Journal of Economic Perspectives* 24(2), 109-128.

Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger (2006): "Identifying Effective Teachers Using Performance on the Job," *Brookings Institution: Hamilton Project Discussion Paper.*

Graham, Bryan S. (2008): "Identifying Social Interactions Through Conditional Variance Restrictions," *Econometrica* 76, 643-660.

Greene, William (2008). "Functional Forms for the Negative Binomial Model for Count Data," *Economics Letters* 99, 585-590.

Gurmu, Shiferaw, Paul Rilstone, and Steven Stern (1999): "Semiparametric Estimation of Count Regression Models," *Journal of Econometrics* 88, 123-150.

Hanushek, Eric A. (1986): "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature* 49(3), 1141-1177.

Hanushek, Eric A., Paul E. Peterson, and Ludger Woessmann (2011): "Teaching Math to the Talented," *Education Next*, 11(1), 11-18.

Hanushek, Eric A. and Ludger Woessmann (2008): "The Role of Cognitive Skills in Economic Development," *Journal of Economic Literature* 46(3), 607-668.

Hoxby, Caroline M. (2000): "Does Competition Among Public Schools Benefit Students and Taxpayers?" *American Economic Review* 90(5), 1209-1238.

Hoxby, Caroline M. (2003): "School Choice and School Productivity: Could School Choice be a Tide That Lifts All Boats?" in *The Economics of School Choice*, ed. Caroline M. Hoxby (Cambridge, MA: National Bureau of Economic Research), 287-342.

Hoxby, Caroline M., Sonali Murarka, and Jenny Kang (2009): "How New York City's Charter Schools Affect Achievement, August 2009 Report." Second report in series. Cambridge, MA: New York City Charter Schools Evaluation Project.

Hoxby, Caroline M. and Christopher Avery (2013): "The Missing 'One-Offs': The Hidden Supply of High-Achieving, Low Income Students," *Brookings Papers on Economic Activity*, 1-66

Jencks, Christopher S. and Marsha D. Brown (1975): "Effects of High Schools on Their Students," *Harvard Educational Review* 45 (3), 273-324.

Joensen, Suanna Schrøter, and Helena Skyt Nielsen (2009): "Is There a Causal Effect of High School Math on Labor Market Outcomes?," *Journal of Human Resources* 44, 171-198.

Karlin, Samuel (1966): *A First Course in Stochastic Processes.* New York: Academic Press.

Krieg, John M. (2008): "Are Students Left Behind? The Distributional Effects of the No Child Left Behind Act," *Education Finance and Policy* 3(2), 250-281.

Krueger, Alan B., and Mikael Lindahl (2001): "Education for Growth: Why and For Whom?," *Journal of Economic Literature* 39 (4), 1101-1136.

Lavergne, Gary, and Bruce Walker (2001): *Developing a Concordance Between the ACT Assessment and the SAT I: Reasoning Test for The University of Texas at Austin.* Austin, TX: University of Texas, available at http://www.utexas.edu/student/admissions/research/ACT-SATconcordance.html.

Levine, Phillip, and David Zimmerman (1995): "The Benefit of Additional High-School Math and Science Classes for Young Men and Women," *Journal of Business and Economic Statistics* 13, 137-149.

Marcon, Eric, and Florence Puech (2003): "Evaluating the Geographic Concentration of Industries using Distance-Based Methods," *Journal of Economic Geography* 3, 409-428.

"Math Teacher Kay Tipton: A True Chalkboard Champion" (2013), http://chalkboardchampions.org/education/math-teacher-kay-tipton-true-chalkboard-champion/, accessed July 28, 2015.

The Mathematical Association of America, American Mathematics Competitions (2007): *58th Annual Summary of High School Results and Awards.*

Neal, Derek and Diane Whitmore Shanzenbach (2010): "Left Behind by Design: Proficiency Counts and Test-Based Accountability," *Review of Economics and Statistics* 92(2), 263-283.

OECD (2010), *PISA 2009 Results: What Students Know and Can Do – Student Performance in Reading, Mathematics and Science (Volume I)*, available at http://dx.doi.org/10.1787/9789264091450-en.

Pallais, Amanda and Sarah Turner (2006): "Opportunities for Low-Income Students at Top Colleges and Universitites: Policy Initiatives and the Distribution of Students," *National Tax Journal* 59(2), 357-386.

Pope, Devin G. and Justin R. Sydnor (2010): "Geographic Variation in the Gender Differences in Test Scores," *Journal of Economic Perspectives* 24 (2), 95-108.

Redd, C. Kalimah (2004): "Honoring a Prime Figure," *Boston Globe* January 11.

Reist, Margaret (2012): "East Math Teacher Retiring; Course Going with Her," *Lincoln Journal Star* January 24.

Rose, Heather and Julian R. Betts (2004): "The Effect of High School Courses on Earnings," *The Review of Economics and Statistics* 86 (2), 497-513.

Rothstein, Jesse (2005): "SAT Scores, High Schools, and Collegiate Performance Predictions, mimeo, University of California, Berkeley.

Schneider, S. (2001). "In Mr. Fischer's Classroom, Math Equals Magic." *USA Today* November 13, 11D.

Skrondal, A, and S. Rabe-Hesketh (2009). "Prediction in Multilevel Generalized Linear Models." *Journal of the Royal Statistical Society.* Series A 172: 659-87.

Solon, Gary, Marianne E. Page and Greg J. Duncan (2000): "Correlations Between Neighboring Children in Their Subsequent Educational Attainment," *The Review of Economics and Statistics* 82, 383-392.

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

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

U.S. Dept. of Education, National Center for Education Statistics: *Common Core of Data: Public School Data, 2005-2006*, available at http://nces.ed.gov/ccd/bat/.

U.S. Dept. of Education, National Center for Education Statistics: *PSS Private School Universe Survey Data, 2005-2006*, available at http://nces.ed.gov/pubsearch/getpubcats.asp?sid=002.

	AMC 12 scores							Prior
	80's	90's	100's	110's	120's	130's	140's	800 SAT
Mean on 1st Math SAT	711	745	773	774	791	793	800	752
% with 800 on 1st SAT	0	19	35	38	65	60	100	15
Sample size	12	32	83	21	32	10	5	—

Table 1: SAT math scores for a sample of students with AMC 12 scores in various ranges

Variable	Mean	St.Dev.
Count AMC > 100	0.80	1.87
Count AMC > 120	0.11	0.49
Count AMC > 100 female	0.12	0.45
Count AMC > 120 female	0.01	0.11
Number of students	1,452.02	819.90
School frac. Asian	0.07	0.12
School frac. black	0.09	0.14
School frac. Hispanic	0.09	0.13
School frac. female	0.49	0.02
Title I school	0.21	0.41
School frac. free lunch	0.15	0.14
log(ZIP median income)	10.84	0.38
Adult frac. BA	0.20	0.09
Adult frac. grad	0.13	0.09
ZIP frac. urban	0.78	0.32
Count perfect SAT/ACT	1.67	3.59
Average SAT/ACT	526.73	41.79
SAT/ACT participation	221.06	153.53

Table 2: School-level summary statistics

	Dependent Variable: Count of high scorers in school					
Variable	AMC12 $\geq$ 100		AMC12 $\geq$ 120	Perfect SAT/ACT		Female AMC12 $\geq$ 100
log(Num. of Students)	1.14	1.13	1.47	1.23	1.22	1.38
	(0.10)	(0.12)	(0.28)	(0.07)	(0.17)	(0.22)
Adult frac. BA	2.50	2.10	6.74	1.34	1.27	3.14
	(0.73)	(0.78)	(1.81)	(0.49)	(0.49)	(1.44)
Adult frac. Grad	7.04	7.32	7.75	5.33	5.29	7.01
	(0.54)	(0.67)	(1.18)	(0.35)	(0.74)	(0.94)
School frac. Asian	1.81	1.80	2.17	2.35	2.36	1.74
	(0.28)	(0.30)	(0.62)	(0.16)	(0.34)	(0.48)
School frac. Black	-0.77	-0.60	-1.44	-1.18	-1.26	-1.19
	(0.43)	(0.44)	(1.34)	(0.32)	(0.36)	(1.00)
School frac. Hisp.	-1.77	-1.77	-3.26	-0.83	-0.86	-1.55
	(0.45)	(0.48)	(1.43)	(0.31)	(0.31)	(0.98)
log(ZIP median income)	-0.70	-0.67	-1.40	-0.18	-0.14	-0.72
	(0.15)	(0.16)	(0.35)	(0.10)	(0.10)	(0.28)
School free lunch frac.	-2.26	-2.47	-3.72	-2.56	-2.39	-2.70
	(0.59)	(0.64)	(1.80)	(0.44)	(0.53)	(1.30)
Title 1 school	-0.04	-0.04	-0.23	0.02	0.01	0.03
	(0.11)	(0.11)	(0.29)	(0.07)	(0.07)	(0.22)
ZIP frac. urban	0.20	0.20	0.60	0.46	0.49	0.14
	(0.24)	(0.25)	(0.78)	(0.18)	(0.18)	(0.55)
School frac. female	-0.30	-0.36	1.86	2.57	2.53	2.77
	(2.24)	(2.26)	(5.94)	(1.53)	(1.56)	(4.69)
log(SAT/ACT participation)				0.95	0.95	
				(0.10)	(0.15)	
Constant	-2.35	-2.65	-1.73	-7.66	-7.98	-7.48
	(2.00)	(2.10)	(4.87)	(1.44)	(1.68)	(4.01)
log-likelihood	-1,899.1	-1,893.6	-493.4	-2,372.5	-2,371.5	-605.6
Pseudo R^2	0.19		0.21	0.27		0.19
Estimation Method	NB	semi-P	NB	NB	semi-P	NB
Estimated Var(u_i)	0.73	0.96	2.18	0.23	0.23	0.95
(Std. Error)	(0.08)	(0.17)	(0.50)	(0.03)	(0.09)	(0.29)
# of obs.	1,984	1,984	1,984	1,984	1,984	1,984
# of high scorers	1,596	1,596	211	3,307	3,307	243

Notes and sources: Results of negative binomial regression and semi-parametric model estimation. Outcomes are counts of high achievers (students scoring more than 100 or 120 on the AMC 12 or with 800(36) on the SAT(ACT) math) in each school. School demographics are from the NCES Common Core of Data for 2005-6; ZIP code demographics are from the 2000 U.S. census. The school sample includes coed, non-charter, non-magnet public schools that offered the 2007 AMC 12.

Table 3: Demographic Predictors of High Math Achievement

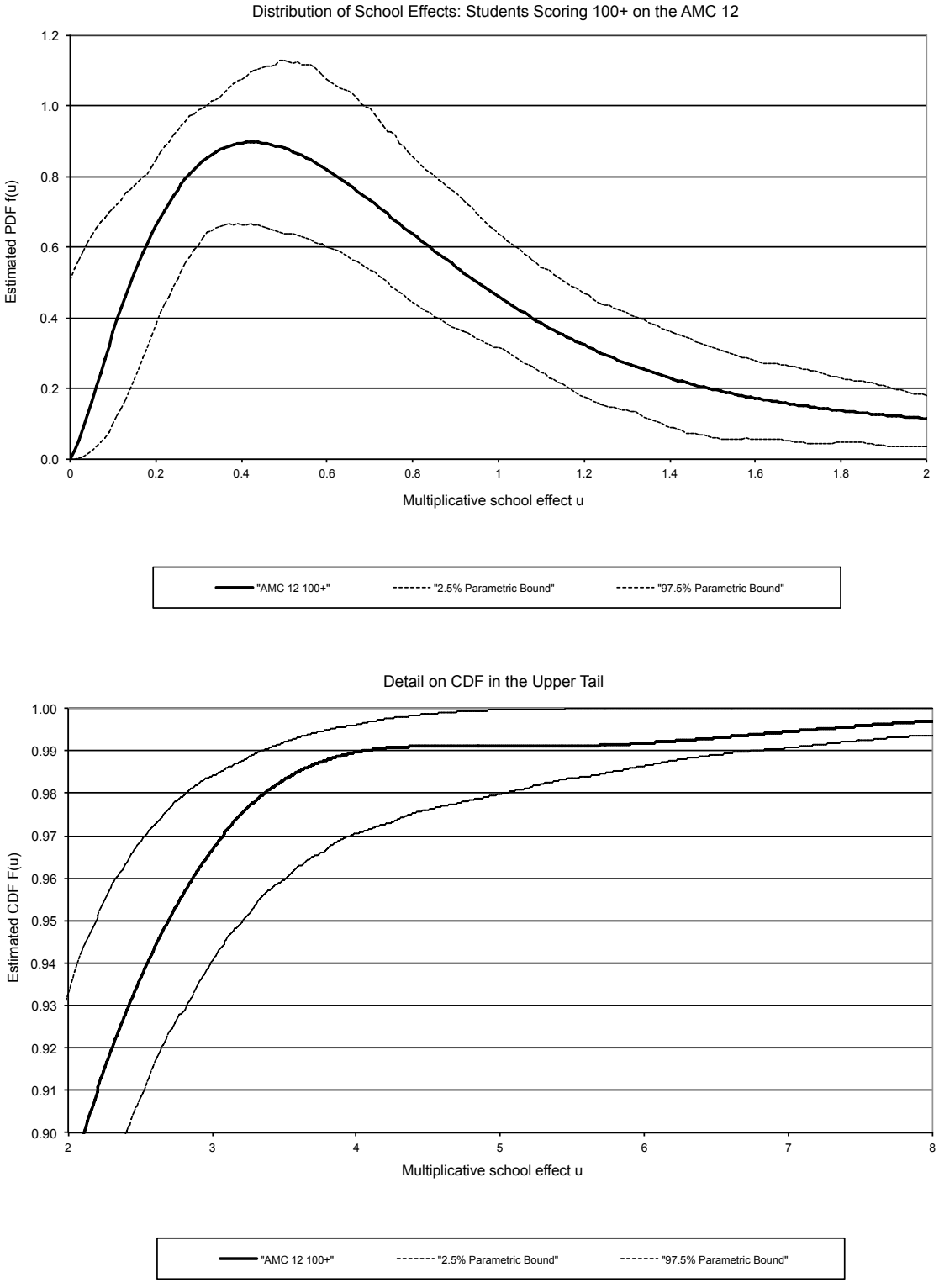

Figure 1: Estimated distribution of school effects: AMC 12 high scorers

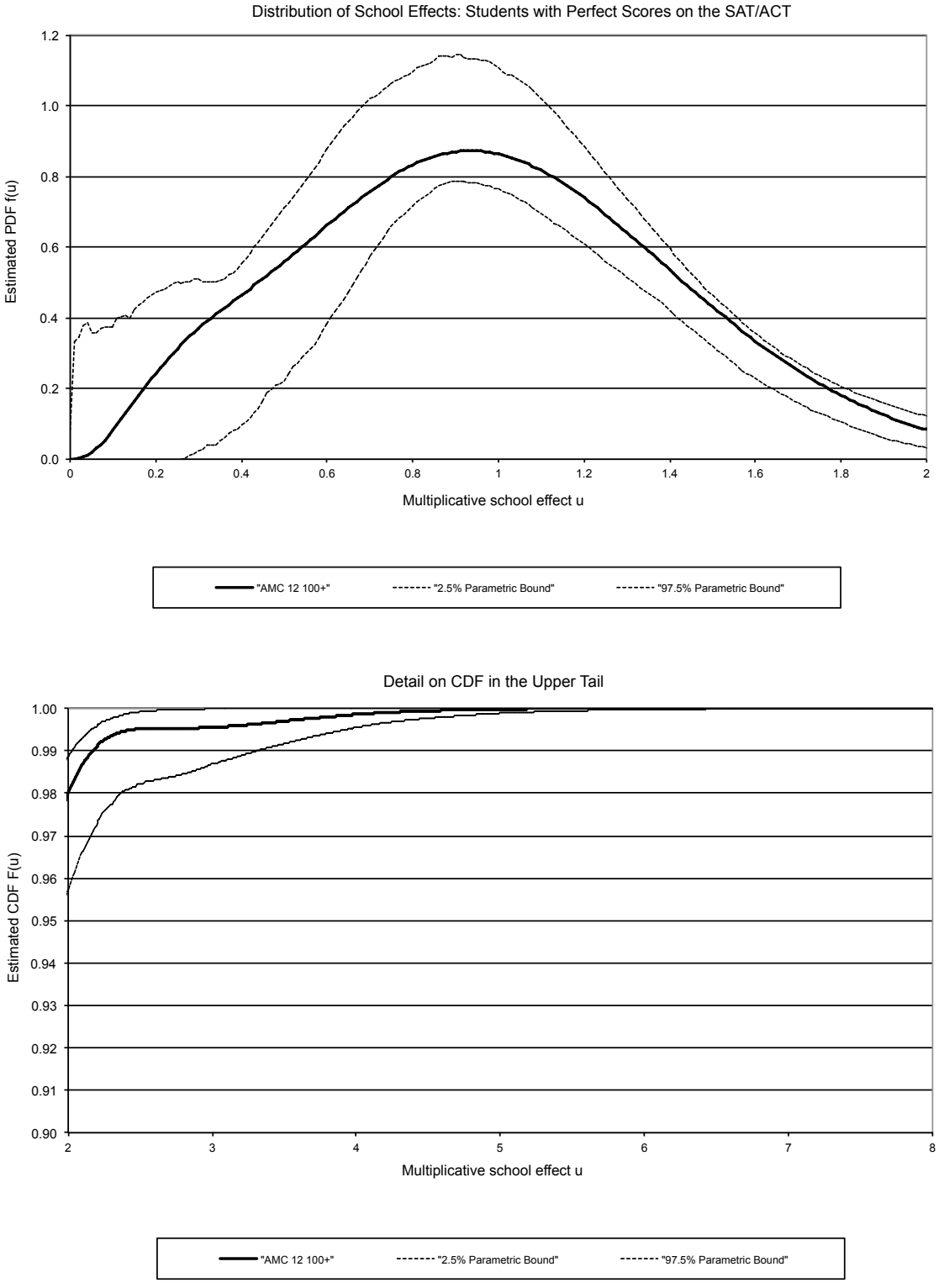

Figure 2: Estimated distribution of school effects: high SAT/ACT scorers

School	Location	AMC > 100 Act.	AMC > 100 Pred.	Post. Mean $\overline{u}_i$	"Star" Teach. Found	Math Club on Wiki	Math Levels Geo/Alg2	Multi-Var. Calc.	Special Features
Vestavia Hills HS	Birmingham, AL	19	1.6	8.9	Y	Y	3/4	N	Yes-1
Terre Haute S. Vigo	Terre Haute, IN	7	0.5	7.6	Y	N	2/2	N	N
Canton HS	Canton, MA	6	0.4	7.2	Y	N	3/3	N	Yes-1
Lassiter HS	Marietta, GA	8	0.9	6.3	Y	Y	3/3	Y	N
Lincoln East HS	Lincoln, NE	8	1.0	5.8	Y	N	3/2	N	N
Westford Academy	Westford, MA	12	1.8	5.7	N	Y	3/4	N	N
Mark Keppel HS	Alhambra, CA	8	1.1	5.6	N	N	3/3	N	N
Revere HS	Richfield, OH	4	0.3	5.3	N	N	2/3	N	N
O'Connor HS	Helotes, TX	6	0.7	5.1	N	N	2/2	Y	N
Grace King HS	Metairie, LA	4	0.4	4.5	N	N	Unk	Unk	Unk
Moses Lake HS	Moses Lake, WA	3	0.2	4.5	N	N	2/2	N	N
Carlmont HS	Belmont, CA	8	1.4	4.1	N	N	3/2	Y	N
Portland HS	Portland, ME	3	0.2	3.8	N	N	3/4	N	N
Sycamore HS	Cincinnati, OH	11	2.1	3.8	Y	Y	4/4	Y	Yes-1
Georgetown HS	Georgetown, TX	3	0.3	3.7	N	N	4/2	N	Yes-1
Oak Ridge HS	Conroe, TX	3	0.3	3.5	N	N	4/4	N	N
Pearland HS	Pearland, TX	4	0.5	3.5	N	N	2/2	N	N
Oak Ridge HS	Oak Ridge, TN	5	0.8	3.4	Y	Y	4/4	Y	Yes-2
Cardozo HS	Bayside, NY	12	2.8	3.1	N	N	3/3	N	Yes-1,2
Manzano HS	Albuquerque, NM	3	0.4	3.0	Y	N	2/2	N	N
Mean for 20 High-Achieving Schools		6.9	0.9	4.9	0.4	0.3	2.9/2.9	0.3	0.3
Mean for 20 Comparison Schools		1.0	0.8	1.0	0.0	0.2	2.2/2.3	0.1	0.0

Table 4: A List of High-Achieving Schools

Appendix

In this appendix we describe how we estimate the distribution of unobserved heterogeneity and present some simulation results on the method.

A.1 Estimation Methodology

Suppose y_i is distributed Poisson (λ_i), where $\lambda_i = e^{z_i\beta}u_i$, z_i is a vector of observable characteristics, and u_i is an unobserved characteristic with a multiplicative effect on the Poisson rate. Assume that the u_i are i.i.d. random variables independent of the z_i with continuous density f on $(0,\infty)$ and $E(u_i) = 1$. We wish to estimate both the coefficients β on the observable characteristics and the distribution f of the unobserved effects.

Our approach is similar to that of Gurmu, Rilstone, and Stone (1999) in that we use a series expansion and exploit known properties of the orthogonal polynomials involved to facilitate maximum likelihood estimation.[50]

Given any function f and any constant α we can write $f(x) = x^\alpha e^{-x}g(x)$. If $g(x)$ is well behaved in the sense that $\int_{x=0}^{\infty} x^\alpha e^{-x}|g(x)|^2dx < \infty$, then $g(x)$ can be represented as a convergent sum

$$g(x) = \sum_{j=0}^{\infty} g_j L_j^{(\alpha)}(x)$$

where $L_j^{(\alpha)}(x)$ is the j^{th} generalized Laguerre polynomial, $L_j^{(\alpha)}(x) \equiv \sum_{i=0}^{j}(-1)^i\binom{j+\alpha}{j-i}\frac{x^i}{i!}$.[51] Expressing the distribution in this way makes it possible to evaluate the likelihood of each outcome without integrating over the unobserved parameter u_i.

Proposition 4 *Consider the model described above. Then,*

$$Pr\{y_i = k|z_i\} = \frac{e^{kz_i\beta}}{\left(e^{z_i\beta}+1\right)^{k+\alpha+1}}\left[\sum_{\ell=0}^{k}\frac{\Gamma(\ell+\alpha+1)}{\ell!}e^{-\ell z_i\beta}(-1)^\ell\binom{k+\alpha}{k-\ell}\left(\sum_{j=\ell}^{\infty}g_j\left(\frac{e^{z_i\beta}}{e^{z_i\beta}+1}\right)^j\binom{j+\alpha}{j-\ell}\right)\right].$$

[50] Our motivation for estimating the distribution as we do rather than directly following Gurmu *et al.* (1999) or Brännäs and Rosenqvist (1994) is that previous Monte Carlo studies have suggested that they may not work well in a situation like ours where we suspect that the distribution is fat-tailed. This is intuitive: Gurmu, *et al.* is based on a flexible estimation the moment-generating function, which is not well defined for a fat-tailed distribution such as the lognormal. Our approach will have some drawbacks relative to that of Gurmu, *et al.* (1999). Most notably, their approach can be applied to any conditional mean function whereas ours works only for conditional mean functions of the form $E(y|X) = e^{X\beta}$. Their expansion is also guaranteed to produce a valid estimated density whereas our estimated densities can take on negative values.

[51] The coefficients g_j are given by

$$g_j = \int_0^\infty \frac{L_j^{(\alpha)}(x)}{\binom{j+\alpha}{j}}g(x)\frac{x^\alpha e^{-x}}{\Gamma(\alpha+1)}dx.$$

When α is not an integer the binomial coefficients are generalized binomial coefficients defined via the gamma function. Note that the parameter α in this section is a completely different parameter from the α we talked about in connection with the negative binomial model. We recognize that this has the potential to cause confusion, but felt we should use α for both parameters because this is how textbook discussions of the negative binomial model and of generalized Laguerre polynomials will refer to the parameters.

The derivation of the formula exploits several properties of the Laguerre polynomials. Details are given in the Online Appendix.

Given the formula above it is natural to estimate the model by maximum likelihood: we simply treat β, α, and the g_j as parameters to be estimated as in a series estimation.[52] For the estimated $f(u)$ to be a valid density with $E(u) = 1$, the estimated parameters $\alpha, g_0, g_1, \ldots, g_N$ must be such that

- $\int_0^\infty u^\alpha e^{-u} \sum_{j=0}^N g_j L_j^{(\alpha)}(u)\ du = 1$; and
- $\int_0^\infty u^{\alpha+1} e^{-u} \sum_{j=0}^N g_j L_j^{(\alpha)}(u)\ du = 1$; and
- $u^\alpha e^{-u} \sum_{j=0}^N g_j L_j^{(\alpha)}(u) \geq 0$ for all $u \in (0, \infty)$.

The first of these conditions holds if and only if $g_0 = 1/\Gamma(\alpha+1)$. We impose this restriction in all of our estimations. Given this restriction, the second condition holds if and only if $g_1 = \frac{\alpha}{\Gamma(\alpha+2)}$. One could impose this restriction, but for finite N it is not necessary for identification. We have chosen instead not to impose it and instead simply renormalize the estimated distribution by dividing the probability density function by the expectation after the estimation stage. We have two motivations for this: first, with the renormalization approach, our model nests the negative binomial, whereas it would not if we had imposed the restriction on g_1; and second, the model seemed to less often produce estimates that ran into the third nonnegativity constraint when we took the renormalization approach. The third constraint is not as easy to express as a parameter restriction, so we do not impose it as a constraint. Instead, we add a penalty function of $-\log\left(\int_0^\infty \max(\hat{f}(x), 0)dx\right)$ to the per-observation likelihood function for parameter values that do not generate valid densities.[53] In practice has the effect of making the estimated densities at most slightly negative.

The function $f(x)$ can in theory be estimated consistently by allowing the number of terms N to grow at an appropriate rate or by choosing it in other other ways like cross-validation. In practice, the number of Laguerre coefficients that can be estimated may be quite limited unless the dataset is very large. It is for this reason that we wrote the density in the form $f(x) = x^\alpha e^{-x} \sum_{j=0}^N g_j L_j^{(\alpha)}(x)$ rather than just as $e^{-x} \sum_{j=0}^N g_j L_j^{(0)}(x)$. When $N = 0$, the α parameter gives the model the ability to fit a range of plausible densities with just a single estimated parameter: the renormalized distribution with parameter α has mean 1 and variance $1/(\alpha+1)$. This allows the model produce an exponential distribution ($\alpha = 0$), unimodal distributions concentrated around one (the distribution is unimodal with mode $\frac{\alpha}{\alpha+1}$ if $\alpha > 0$), and distributions with more weight on extreme u's than the exponential ($\alpha \in (-1, 0)$).[54] The objective function is not globally concave, so we ran our

[52] Finite sums $g^N(x) \equiv \sum_{j=0}^N g_j L_j^{(\alpha)}(x)$ will approximate the true distribution as $N \to \infty$. Defining $\|g - g^N\| \equiv \int_{x=0}^\infty (g(x) - g^N(x))^2 \frac{x^\alpha e^{-x}}{\Gamma(\alpha+1)} dx$ we have $\|g - g^N\| \leq \sum_{j=N+1}^\infty \binom{j+\alpha}{j} g_j^2$.

[53] The motivation for the form of the penalty is that, given a weighting function $\hat{f}(x)$ that is not everywhere nonnegative, one can define a nonnegative measure by setting $\tilde{f}(x) = \max(\hat{f}(x), 0) / \int_0^\infty \max(\hat{f}(x), 0)dx$. The likelihood minus the penalty function is a lower bound to the likelihood that would be obtained from the nonnegative density $\tilde{f}(x)$.

[54] The pure Poisson model with no unobserved heterogeneity is obtained as a special case as $\alpha \to \infty$. The negative binomial is also a special case if we adopt the normalization strategy of estimating the distribution

estimation routines from a large number of starting values for α and the g_i.[55]

However, the normalization is done one can calculate the variance of the estimated distribution using Laguerre polynomial identities and get an estimated variance that can be compared to the α in the negative binomial model. The normalization strategy we use is to first estimate without the restriction on g_1. This gives a distribution of shocks v that have $E(v) = (1+\alpha) - g_1\Gamma(\alpha+2)$. We get a distribution u with mean 1 by defining $u = v/((1+\alpha) - g_1\Gamma(\alpha+2))$. The variance of the resulting random variable is

$$\text{Var}(u) = E(u^2) - E(u)^2 = \frac{E(v^2)}{E(v)^2} - 1 = \frac{\Gamma(\alpha+3)g_2 - 2(\alpha+2)\Gamma(\alpha+2)g_1 + (\alpha+1)(\alpha+2)}{((1+\alpha) - g_1\Gamma(\alpha+2))^2} - 1$$

If we were to impose the restriction $g_1 = \frac{\alpha}{\Gamma(\alpha+2)}$, then the formula simplifies substantially to

$$\text{Var}(u) = \Gamma(\alpha+3)g_2 - \alpha^2 - \alpha + 1.$$

A.2 Simulation Results

Our primary motivation for estimating the model as described above instead of directly following previous approaches is that simulations have suggested that previous approaches may not work well in practice when the distribution of unobserved heterogeneity is fat tailed.[56] To assess how our method might work in practice and how many terms N one might want to include in the series expansion we also conducted simulation experiments described in the Online Appendix using exponential, log-normal, and uniform distributions for the unobserved heterogeneity. A very rough summary is that our approach seems to work reasonably well in the exponential and log-normal cases. Estimating the upper tail is easier than estimating the density at low values of u: it is inherently very difficult to distinguish whether a school is producing 0.1 or 0.01 high-achieving students per year. The simulations also suggest that including $N = 4$ terms in the series expansion may a good choice for balancing flexibility vs. overfitting given the number of observations in our dataset and the magnitudes of the counts. In our empirical analyses, we will generally present estimates that use $N = 4$ terms in the series expansion.

---

without the restriction on g_1 and then rescaling the estimated distribution so that it does have expectation 1 by dividng by its mean.

[55] For starting values, we used the coefficients from the negative binomial regressions for β_{start}; for α_{start} and $g_{i,start}$, we tried the mean estimates from the uniform, exponential, and lognormal Monte Carlo exercise, as well as setting each $g_{i,start} = 0$ and varying α_{start} between -1 and 2 in increments of 0.01. The latter approach (varying α_{start} with each $g_{i,start} = 0$) generally resulted in the largest log-likelihood.

[56] See Gurmu, *et al.* (1999) p. 141.

A.3 Sample Test Questions

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

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

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

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

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

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

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

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

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

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

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

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

18. The polynomial $f(x) = x^4 + ax^3 + bx^2 + cx + d$ has real coefficients, and $f(2i) = f(2 + i) = 0$. What is $a + b + c + d$?

    (A) 0 **(B)** 1 **(C)** 4 **(D)** 9 **(E)** 16

19. Triangles ABC and ADE have areas 2007 and 7002, respectively, with $B = (0, 0)$, $C = (223, 0)$, $D = (680, 380)$, and $E = (689, 389)$. What is the sum of all possible x-coordinates of A?

    (A) 282 **(B)** 300 **(C)** 600 **(D)** 900 **(E)** 1200

20. Corners are sliced off a unit cube so that the six faces each become regular octagons. What is the total volume of the removed tetrahedra?

    (A) $\frac{5\sqrt{2} - 7}{3}$ **(B)** $\frac{10 - 7\sqrt{2}}{3}$ **(C)** $\frac{3 - 2\sqrt{2}}{3}$ **(D)** $\frac{8\sqrt{2} - 11}{3}$

    (E) $\frac{6 - 4\sqrt{2}}{3}$

Figure 3: Questions 13 through 20 from the 2007 AMC 12A

3. Let $P(x) = a_n x^n + a_{n-1} x^{n-1} + ... + a_1 x + a_0$ be a polynomial with integer coefficients. Suppose there exist distinct integers a, b, c, d such that $P(a) = P(b) = P(c) = P(d) = 4$. Prove that there exists no integer m such that $P(m) = 7$.

4. The lengths of two sides of an equilateral triangle are doubled, creating an isosceles triangle, as shown in the diagram at the right. These two longer sides are doubled again creating a third isosceles triangle (all three triangles having the same base). This process is continued indefinitely. If the measure of the vertex angle of each triangle is represented by $A_1, A_2, A_3, ...$, determine, with proof, the value of $(1 - \cos A_1) + (1 - \cos A_2) + (1 - \cos A_3) + \ldots$.

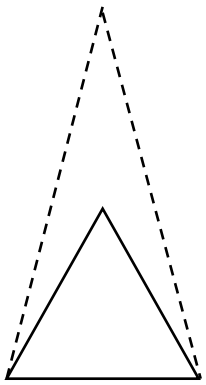

Figure 4: Questions 3 and 4 from the 2007 Kennesaw (GA) State University HS Mathematics Competition

Online Appendix

1 Appendix Tables

Count	Actual	Predicted # of schools		
		Poisson	NB	Semi-P
0	1,311	975	1,311	1,309
1	357	608	355	362
2	135	267	140	139
3	64	94	67	65
4	39	29	36	35
5	18	8	22	21
6	15	2	14	13
7	17	0	9	9
8	10	0	7	6
9	4	0	5	5
10	2	0	4	4
11	3	0	3	3
12	3	0	2	2
13	1	0	2	2
14	0	0	1	1
15	0	0	1	1
16	0	0	1	1
17	0	0	1	1
18	1	0	1	1
19	1	0	1	1
20+	3	0	4	4
log-likelihood		-2,063.6	-1,899.1	-1,893.6
χ^2		841.5E+08	15.5	16.5
p-value		0.000	0.745	0.688

Table 5: Actual vs. predicted distribution of counts of high-scorers across schools

2 Proofs

Proof of Proposition 1

It is a standard that under model 1 $Y_i \sim NB\left(\frac{1}{\alpha}, \frac{\alpha e^{X_i\beta}}{1+\alpha e^{X_i\beta}}\right)$ and under model 2 $Y_i \sim NB\left(\frac{\lambda(X_i)}{g(X_i)}, 1-e^{-g(X_i)}\right)$. (See Boswell and Patil (1970) or Karlin (1966, p. 345).) Hence,

the distributions in the two models are identical if

$$\begin{aligned} \frac{1}{\alpha} &= \frac{\lambda(X_i)}{g(X_i)}; \text{ and} \\ \frac{\alpha e^{X_i\beta}}{1+\alpha e^{X_i\beta}} &= 1-e^{-g(X_i)}. \end{aligned}$$

The first holds for all X_i if $g(X_i) = \alpha\lambda(X_i)$. The second then holds if

$$\frac{1}{1+\alpha e^{X_i\beta}} = e^{-g(X_i)},$$

which holds for $g(X_i) = \log\left(1+\alpha e^{X_i\beta}\right)$. □

Proof of Proposition 2

Applying the result of Proposition 1 to the outcome of this model conditional on u_i we see that the conditional distribution is $NB\left(\frac{1}{\alpha_p}, \frac{\alpha_p e^{X_i\beta}u_i}{1+\alpha_p e^{X_i\beta}u_i}\right)$. The mean and variance of a $NB(r,p)$ distribution are $E(Y) = \frac{rp}{1-p}$ and $\text{Var}(Y) = \frac{rp}{(1-p)^2} = \frac{E(Y)}{1-p}$.This gives $E(Y_i|X_i,u_i) = e^{X_i\beta}u_i$ and $\text{Var}(Y_i|X_i,u_i) = E(Y_i|X_i,u_i) + \alpha_p E(Y_i|X_i,u_i)^2$. The result on the expection of $Y_i|X_i$ follows from iterated expectations:

$$E(Y_i|X_i) = E_{u_i}\left(E(Y_i|X_i,u_i)\right) = E_{u_i}\left(e^{X_i\beta}u_i\right) = e^{X_i\beta}.$$

And the formula for the variance follows from the conditional variance formula:

$$\begin{aligned} \text{Var}(Y_i|X_i) &= E_{u_i}\text{Var}(Y_i|X_i,u_i) + \text{Var}_{u_i}E(Y_i|X_i,u_i) \\ &= E_{u_i}\left(e^{X_i\beta}u_i + \alpha_p e^{2X_i\beta}u_i^2\right) + \text{Var}_{u_i}\left(e^{X_i\beta}u_i\right) \\ &= e^{X_i\beta}Eu_i + \alpha_p e^{2X_i\beta}\left(\text{Var}(u_i) + (Eu_i)^2\right) + e^{2X_i\beta}\text{Var}(u_i) \\ &= e^{X_i\beta} + \alpha_p e^{2X_i\beta}(\alpha_u+1) + e^{2X_i\beta}\alpha_u \\ &= e^{X_i\beta} + e^{2X_i\beta}(\alpha_p\alpha_u + \alpha_p + \alpha_u). \quad \square \end{aligned}$$

Proof of Proposition 3

Suppose the Y_{it} are generated as described. Then using Proposition 2 and iterated expectations over t we have

$$E(Y_{it}|X_i) = \frac{1}{2}E(Y_{i1}|X_i) + \frac{1}{2}E(Y_{i2}|X_i) = \frac{1}{2}e^{X_i\beta} + \frac{1}{2}e^{X_i\beta} = e^{X_i\beta}.$$

The conditional variance formula gives $\text{Var}(Y_{it}|X_i) = E_t\text{Var}(Y_{it}|X_i,t) + \text{Var}_t(E(Y_{it}|X_i))$. The first term on the RHS of this expression is just the variance in the single period model given by Proposition 2, and the second term is zero, so we find

$$\text{Var}(Y_{it}|X_i) = e^{X_i\beta} + e^{2X_i\beta}(\alpha_p\alpha_u + \alpha_p + \alpha_u).$$

The mean of $\overline{Y}_i|X_i$ follows from an identical calculation:

$$E(\overline{Y}_i|X_i) = E(Y_{i1}+Y_{i2}) = E(Y_{i1}|X_i) + E(Y_{i2}|X_i) = 2e^{X_i\beta}.$$

The variance is is a little more complicated. We have

$$\text{Var}(\overline{Y}_i|X_i) = \text{Var}\left(Y_{i1} + Y_{i2}\right) = \text{Var}(Y_{i1}|X_i) + \text{Var}(Y_{i2}|X_i) + 2\text{Cov}(Y_{i1}, Y_{i2}|X_i).$$

To find the covariance we condition on u_i and use the fact that Y_{i1} and Y_{i2} are conditionally independent given X_i and u_i:

$$\begin{aligned}
\text{Cov}(Y_{i1}, Y_{i2}|X_i) &= E(Y_{i1}Y_{i2}|X_i) - E(Y_{i1}|X_i)E(Y_{i2}|X_i) \\
&= E_{u_i}\left(E(Y_{i1}Y_{i2}|X_i, u_i)\right) - e^{2X_i\beta} \\
&= E_{u_i}\left(E(Y_{i1}|X_i, u_i)E(Y_{i2}|X_i, u_i)\right) - e^{2X_i\beta} \\
&= E_{u_i}\left(e^{X_i\beta}u_i e^{X_i\beta}u_i\right) - e^{2X_i\beta} \\
&= e^{2X_i\beta}\left(E_{u_i}(u_i^2 t) - 1\right) = e^{2X_i\beta}\alpha_u
\end{aligned}$$

Plugging back into the formula for the variance we find

$$\begin{aligned}
\text{Var}(\overline{Y}_i|X_i) &= 2\left(e^{X_i\beta} + e^{2X_i\beta}(\alpha_u + \alpha_p + \alpha_u\alpha_p)\right) + 2e^{2X_i\beta}\alpha_u \\
&= \left(2e^{X_i\beta}\right) + \left(2e^{X_i\beta}\right)^2\left(\alpha_u + \frac{1}{2}\alpha_p + \frac{1}{2}\alpha_p\alpha_u\right)
\end{aligned}$$

Using these formulas we will have $\text{Var}(Y_{it}|X_i) = E(Y_{it}|X_i) + \alpha E(Y_{it}|X_i)^2$ and $\text{Var}(\overline{Y}_i|X_i) = E(\overline{Y}_i|X_i) + \overline{\alpha}E(\overline{Y}_i|X_i)^2$ if and only if two conditions hold:

$$\begin{aligned}
\alpha &= \alpha_u + \alpha_p + \alpha_u\alpha_p; \quad \text{and} \\
\overline{\alpha} &= \alpha_u + \frac{1}{2}\alpha_p + \frac{1}{2}\alpha_u\alpha_p.
\end{aligned}$$

The first equation can hold for nonnegative (α_u, α_p) only if $\alpha_u \in [0, \alpha]$. Given any such α_u the first equation will hold for an unique α_p: $\alpha_p(\alpha_u) \equiv \frac{\alpha - \alpha_u}{1 + \alpha_u}$. Given this value for α_p we have $\alpha_p + \alpha_p\alpha_u = \alpha - \alpha_u$ so the second equation becomes $\overline{\alpha} = \alpha_u + \frac{1}{2}(\alpha - \alpha_u)$ which is true for $\alpha_u = 2\overline{\alpha} - \alpha$. The formula for α_p follows by substitution. □

Proof of Proposition 4

Let the density $f(x)$ be represented as $f(x) = x^\alpha e^{-x}\sum_{j=0}^\infty g_j L_j^{(\alpha)}(x)$.The distribution of y_i is then described by

$$\begin{aligned}
\Pr\{y_i = k|\tilde{z}_i\} &= \int_0^\infty e^{-\left(e^{\tilde{z}_i\beta}u_i\right)}\frac{\left(e^{\tilde{z}_i\beta}u_i\right)^k}{k!}u_i^\alpha e^{-u_i}\left(\sum_{j=0}^\infty g_j L_j^{(\alpha)}(u_i)\right)\, du_i \\
&= \int_0^\infty e^{-\left(e^{\tilde{z}_i\beta}+1\right)u_i}\frac{\left(e^{\tilde{z}_i\beta}u_i\right)^k}{k!}u_i^\alpha\left(\sum_{j=0}^\infty g_j L_j^{(\alpha)}(u_i)\right)\, du_i.
\end{aligned}$$

Let $z_i = \left(e^{\tilde{z}_i\beta} + 1\right)u_i$, so $dz_i = \left(e^{\tilde{z}_i\beta} + 1\right)du_i$. Then

$$\begin{aligned}
Pr\{y_i = k|\tilde{z}_i\} &= \int_0^\infty e^{-z_i}\frac{z_i^k}{k!}\left[\frac{e^{\tilde{z}_i\beta}}{e^{\tilde{z}_i\beta}+1}\right]^k\frac{1}{\left(e^{\tilde{z}_i\beta}+1\right)^\alpha}z_i^\alpha\left(\sum_{j=0}^\infty g_j L_j^{(\alpha)}\left(\frac{z_i}{e^{\tilde{z}_i\beta}+1}\right)\right)\, \frac{dz_i}{e^{\tilde{z}_i\beta}+1} \\
&= \frac{e^{k\tilde{z}_i\beta}}{\left(e^{\tilde{z}_i\beta}+1\right)^{k+\alpha+2}}\int_0^\infty e^{-z_i}\frac{z_i^k}{k!}z_i^{\alpha+1}\left(\sum_{j=0}^\infty g_j L_j^{(\alpha)}\left(\frac{z_i}{e^{\tilde{z}_i\beta}+1}\right)\right)\, dz_i.
\end{aligned}$$

To simplify we use two well-known identities: the monomial formula for Laguerre polynomials,

$$\frac{u_i^k}{k!} = \sum_{l=0}^{k} (-1)^l \left(\begin{array}{c} k+\alpha \\ k-l \end{array} \right) L_l^{(\alpha)}(u_i),$$

and the series expansion

$$L_j^{(\alpha)}\left(\frac{u_i}{1+\gamma}\right) = \frac{1}{(1+\gamma)^j} \sum_{l=0}^{j} \gamma^{j-l} \left(\begin{array}{c} j+\alpha \\ j-l \end{array} \right) L_l^{(\alpha)}(u_i).$$

The former implies that

$$\frac{z_i^k}{k!} = \sum_{l=0}^{k} (-1)^l \left(\begin{array}{c} k+\alpha \\ k-l \end{array} \right) L_l^{(\alpha)}(z_i)$$

and the latter implies that

$$L_j^{(\alpha)}\left(\frac{z_i}{e^{\tilde{z}_i\beta}+1}\right) = \frac{1}{(e^{\tilde{z}_i\beta}+1)^j} \sum_{l=0}^{j} e^{(j-l)\tilde{z}_i\beta} \left(\begin{array}{c} j+\alpha \\ j-l \end{array} \right) L_l^{(\alpha)}(z_i).$$

Substituting these formulas into the formula for y_i gives

$$\begin{aligned} Pr\{y_i = k|\tilde{z}_i\} &= \frac{e^{k\tilde{z}_i\beta}}{(e^{\tilde{z}_i\beta}+1)^{k+\alpha+2}} \int_0^\infty z_i^{\alpha+1} e^{-z_i} \left(\sum_{l=0}^{k} (-1)^l \left(\begin{array}{c} k+\alpha \\ k-l \end{array} \right) L_l^{(\alpha)}(z_i) \right) \\ &\quad \left(\sum_{j=0}^{\infty} g_j \frac{1}{(e^{\tilde{z}_i\beta}+1)^j} \sum_{l=0}^{j} e^{(j-l)\tilde{z}_i\beta} \left(\begin{array}{c} j+\alpha \\ j-l \end{array} \right) L_l^{(\alpha)}(z_i) \right) dz_i. \end{aligned}$$

Laguerre polynomials are orthogonal with

$$\int_0^\infty z_i^\alpha e^{-z_i} L_n^{(\alpha)}(z_i) L_m^{(\alpha)}(z_i)\, dz_i = \left\{ \begin{array}{ll} 0 & \text{if} \quad m \neq n \\ \frac{\Gamma(n+\alpha+1)}{n!} & \text{if} \quad m = n \end{array} \right.$$

Using this, the the formula for y_i simplifies to

$$\begin{aligned} Pr\{y_i = k|\tilde{z}_i\} &= \frac{e^{k\tilde{z}_i\beta}}{\left(e^{\tilde{z}_i\beta}+1\right)^{k+\alpha+1}} \left(\sum_{l=0}^{k} (-1)^l \left(\begin{array}{c} k+\alpha \\ k-l \end{array} \right) \frac{\Gamma(l+\alpha+1)}{l!} \right) \left(\sum_{j=l}^{\infty} g_j \frac{1}{\left(e^{\tilde{z}_i\beta}+1\right)^j} e^{(j-l)\tilde{z}_i\beta} \left(\begin{array}{c} j+\alpha \\ j-l \end{array} \right) \right) \\ &= \frac{e^{k\tilde{z}_i\beta}}{\left(e^{\tilde{z}_i\beta}+1\right)^{k+\alpha+1}} \left[\sum_{l=0}^{k} \frac{\Gamma(l+\alpha+1)}{l!} e^{-l\tilde{z}_i\beta} (-1)^l \left(\begin{array}{c} k+\alpha \\ k-l \end{array} \right) \left(\sum_{j=l}^{\infty} g_j \left(\frac{e^{\tilde{z}_i\beta}}{e^{\tilde{z}_i\beta}+1} \right)^j \left(\begin{array}{c} j+\alpha \\ j-l \end{array} \right) \right) \right] \end{aligned}$$

This completes the proof. □

The conditions given in he text for the u to be a valid density, for $E(u) = 1$, and the expression for the $\text{Var}(u)$ can be derived by applying the formula

$$\int_0^\infty z_i^\alpha e^{-z_i} L_n^{(\alpha)}(z_i) L_m^{(\alpha)}(z_i)\, dz_i = \left\{ \begin{array}{ll} 0 & \text{if} \quad m \neq n \\ \frac{\Gamma(n+\alpha+1)}{n!} & \text{if} \quad m = n \end{array} \right.$$

with $n = 0$ and $m = 1, 2, 3$ using $L_0^{(\alpha)}(x) = 1$, $L_1^{(\alpha)}(x) = -x + (1 + \alpha)$, and $L_2^{(\alpha)}(x) = \frac{x^2}{2} - (\alpha + 2)x + \frac{(\alpha+1)(\alpha+2)}{2}$.

3 Bootstrap Procedure

We obtained standard errors for our semiparametric estimates and confidence bands for the distribution of unobserved heterogeneity using both parametric and nonparametric bootstrapping procedures. In each iteration j of the bootstrap, we generate a simulated dataset $\{\tilde{y}_{ij}, \tilde{z}_{ij}\}_{i=1}^{1,984}$, then estimate the parameters $\tilde{\alpha}_j, \tilde{g}_{j1}, ..., \tilde{g}_{jN}, \tilde{\beta}_j$ using the semiparametric estimation procedure described in Section 4. Standard errors are calculated as the standard deviation of each estimated parameter across $1,000$ simulations. For example, the standard error of $\hat{\alpha}$ is calculated as

$$SE(\hat{\alpha}) = \sqrt{\frac{\sum_{j=1}^{1000} (\tilde{\alpha}_j - \hat{\alpha})^2}{1000}}.$$

Another functional of interest is a 95% confidence band on the estimated density and CDF of unobserved heterogeneity. For each $u \in (0, \infty)$ and for each simulation j of the bootstrap, we calculate the density $\tilde{f}_j$ and CDF $\tilde{F}_j$ as those generated by the parameter vector $\tilde{\alpha}_j, \tilde{g}_{j1}, ..., \tilde{g}_{jN}, \tilde{\beta}_j$. Denote as $\tilde{f}_p(u)$ the p^{th} percentile of $\tilde{f}(u)$ across 1,000 simulations; then the 95% confidence band for $\hat{f}(u)$ is $\left(\tilde{f}_{2.5}(u), \tilde{f}_{97.5}(u)\right)$. The confidence band for $\hat{F}$ is calculated similarly. Confidence bands for $u \in (0, 3)$ and $u \in (3, 10)$ are shown in Section 5 for the production of AMC high-scorers and in Section 6 for the production of SAT high-scorers.

In each simulation of the parametric bootstrap, we use the parameter estimates obtained using our semiparametric procedure to generate simulated outcomes. First, we draw a random sample $\tilde{\mathbf{z}}_j$ of size 1,984 (with replacement) from the set of covariates $\mathbf{z}$ listed in Table 3. We also draw a random sample $\tilde{\mathbf{u}}_j$ of size 1,984 from the CDF $\hat{F}$, which we estimated using the procedure in Section 4 on the true dataset. For each $i = 1, ..., 1,984$, we then generate $\lambda_{ji} = e^{\tilde{z}_{ji}\hat{\beta}}\tilde{u}_{ji}$ and draw $\tilde{y}_{ji}^P$ from a Poisson distribution with rate parameter λ_{ji}. Finally, we estimate $\tilde{\alpha}_j^P, \tilde{g}_{j1}^P, ..., \tilde{g}_{jN}^P, \tilde{\beta}_j^P$ on the simulated dataset $(\tilde{\mathbf{y}}_j^P, \tilde{\mathbf{z}}_j)$.

The nonparametric bootstrap proceeds similarly, except that we use the empirical distribution of $\mathbf{y}$ rather than the estimated theoretical distribution of $\mathbf{y}$. That is, for each simulation, we draw a random sample $(\tilde{\mathbf{y}}_j^{NP}, \tilde{\mathbf{z}}_j)$ of size $1,984$ (with replacement) from the set of outcomes $\mathbf{y}$ and covariates $\mathbf{z}$, then estimate $\tilde{\alpha}_j^{NP}, \tilde{g}_{j1}^{NP}, ..., \tilde{g}_{jN}^{NP}, \tilde{\beta}_j^{NP}$ on the simulated dataset $(\tilde{\mathbf{y}}_j^{NP}, \tilde{\mathbf{z}}_j)$. As in the semiparametric estimation on our full sample, the results of each bootstrap estimation may depend on the starting values chosen; in our results, we present those estimates for which the likelihood is highest after trying numerous starting values.[57] We begin each bootstrap by running a trial bootstrap of 20 simulations for several candidate starting values: those resulting in the highest likelihood in the full sample

[57] In practice, we used β starting values from either a Poisson or negative binomial regression, along with one of two potential sets of starting values for our parameters $\alpha, g_1, ..., g_N$. The first set of parameters we tried was the best-fit parameters of the candidate distributions described in Appendix A.2, so that the optimization would be allowed to converge to a number of differently-shaped distributions. We also tried setting each $g_i = 0$ and varying α between -0.9 and 2. The latter approach often yielded the highest likelihood.

estimation and the center of each range of starting values for which the resulting likelihood is close to that of the best starting values. We then use the values that provide the highest average log-likelihood in the trial bootstrap as the starting values in the full bootstrap.

If our model is specified correctly, then the parametric bootstrap is more efficient; if the model is misspecified, then the nonparametric bootstrap will be more appropriate. See Efron and Tibshirani (1993) for a discussion. In our application, neither procedure provides smaller or larger standard errors or confidence bands across all parameters or outcomes, but parametric standard errors are often slightly smaller, and parametric bands are often slightly narrower and smoother. In the body of the paper, we present the results of the parametric bootstrap, but our interpretation of the results is unaffected by the choice of bootstrap procedure.

4 Simulations

The simulations implemented our estimation procedure on datasets created by drawing each z_i from a uniform distribution with support $[0, 1]$; drawing each u_i from the desired error distribution; forming $\lambda_i = e^{z_i\beta}u_i$, where $\beta = [-4.27, 1, 1, 1, 0.1, 0.1, 0.2]$; and drawing y_i from a Poisson distribution with rate parameter λ_i. Each simulated variable included $2,500$ observations. The distributions of the simulated covariates and the values for β were chosen so that the mean and variance of the simulated $e^{z_i\beta}$ would roughly match the mean and variance of the fitted values in a negative binomial regression of the count of AMC 12 high-scorers on school-level covariates. The u_i were chosen from one of three distributions depending on the simulation: an exponential distribution with mean and standard deviation 1, a lognormal distribution with mean 1 and variance $\frac{1}{3}$, and a uniform distribution on $[0, 2]$. The motivation for these choices was to demonstrate the performance of our procedure for a diverse set of underlying distributions: the exponential distribution is within the class of models being estimated even if $N = 0$, the lognormal distribution cannot be fit perfectly with a finite N and has a thicker upper tail, and the uniform distribution is a more challenging distribution to reproduce with a series expansion. We estimated the model using $N = 0, 2, 4, 6$, and 8 terms.[58]

The estimated coefficients $\hat{\beta}$ on the observed characteristics are fairly precise and show almost no bias. Table 6 presents some summary statistics on the estimates for simulations with $N = 8$ Laguerre polynomials.[59] The first column lists the true values for the coefficients on each simulated covariate. The next three columns list the mean and standard deviation (in parentheses) of the estimates across the 1000 simulated datasets for each simulated distribution. There are no notable differences across heterogeneity distributions in the consistency or precision of estimated $\hat{\beta}$'s.

Table 7 provides some statistics on how well the model was able to estimate the distribution of unobserved heterogeneity. The rows correspond to the distribution from which the u's were drawn. The columns correspond to the number N of Laguerre polynomials used in the estimations. The metric used to measure performance is integrated squared error (ISE) – if the estimated density function from simulation run i is $\hat{f}_i(x)$, where the

[58] For these estimations we did not restrict g_1 to be $\alpha/\Gamma(\alpha + 2)$ and instead ensured that the estimated distributions have mean 1 by rescaling the preliminary estimates by dividing by the mean.

[59] Summary statistics for estimates of $\hat{\beta}$ using $N = 0, 2, 4, 6$ are similar.

Variable	True Coeffs.	Mean and SD of estimated coefficients		
		Exponential u	Lognormal u	Uniform u
Constant	-4.270	-4.2690	-4.2651	-4.2777
		(0.1536)	(0.1571)	(0.1109)
z_1	1.000	0.9971	0.9977	0.9984
		(0.1055)	(0.0593)	(0.0760)
z_2	1.000	1.0010	1.0010	1.0026
		(0.0537)	(0.0424)	(0.0401)
z_3	1.000	0.9995	0.9991	1.0019
		(0.0371)	(0.0377)	(0.0269)
z_4	0.100	0.0994	0.0993	0.0998
		(0.0271)	(0.0154)	(0.0190)
z_5	0.100	0.0997	0.0996	0.1011
		(0.0216)	(0.0127)	(0.0151)
z_6	0.200	0.1996	0.1994	0.2003
		(0.0184)	(0.0125)	(0.0132)

Notes: True and estimated coefficients from semi-parametric model estimation using simulated data, varying the distribution of underlying heterogeneity. Results displayed for the exponential (1) distribution, the lognormal $(1, \frac{1}{3})$ distribution, and the uniform $[0, 2]$ distribution with 2,500 simulated observations. Mean estimates across 1,000 simulated datasets shown; standard deviations in parentheses.

Table 6: Estimated coefficients on observed characteristics in simulations

true data generation process has unobserved heterogeneity from distribution $f(x)$, the ISE of that estimated density is $\int_0^\infty (\hat{f}_i(x) - f(x))^2 dx$. The values in Table 7 are median ISE across 1,000 simulation runs.

	Median ISE for various models				
True distribution of u	$N = 0$	$N = 2$	$N = 4$	$N = 6$	$N = 8$
Exponential	0.0010	0.0045	0.0140	0.0201	0.0243
Lognormal	0.0133	0.0115	0.0191	0.0148	0.0167
Uniform $[0, 2]$	0.1055	0.1449	0.0833	0.0795	0.1009

Notes: Median integrated squared error of estimated distributions from semi-parametric model estimation using simulated data, varying the distribution of underlying heterogeneity. Results displayed for the exponential (1) distribution, the lognormal $(1, \frac{1}{3})$ distribution, and the uniform $[0, 2]$ distribution with 2,500 simulated observations. Median ISE across 1,000 simulated datasets shown, varying the number of Laguerre polynomials.

Table 7: Goodness of fit of estimated distributions of unobserved heterogeneity in simulations: median MISE for various models and true distributions

The exponential model fits fairly well for all N. As one would expect, the $N = 0$ fit is best: the true model is in the $N = 0$ class and estimating additional unnecessary parameters just increases the scope for overfitting. The fit worsens gradually as N increases, but never becomes terrible; at $N = 8$, the worst fit, the median ISE is 0.024. To get a feel for the magnitudes, the MISE would be 0.02 if the density of an exponential distribution were over- or under- estimated by 10% at every value of u. Note also that the exponential distribution with mean 1 is the gamma distribution involved in the Poisson-gamma justification for the negative binomial when $\alpha = 1$. Hence, the estimates of this model can provide a sense for how well our semiparametric model will estimate the distribution of underlying heterogeneity in a case where the negative binomial is correctly specified.

The lognormal distribution does not fit as well when $N = 0$. This should be expected: the lognormal is not a member of the parametric family we are estimating and indeed no matter what α is estimated the ISE cannot possibly be below 0.0107. Larger N make it theoretically possible to fit the distribution much better (the parameter vectors that give distributions closest to the true lognormal have ISEs of 0.00756, 0.00210, 0.00014, and 0.00002 for $N = 2, 4, 6$, and 8 respectively), but again there is the offsetting effect that there is more scope for overfitting. The tradeoff between the two effects results in fairly similar fits across the range of N. The median ISE is smallest for the $N = 2$ model.

The fits to the uniform distribution are much worse. Here, there is no parameter combination that produces a very good fit when N is small, and overfitting becomes a concern when N is large.[60] The best fit is obtained for $N = 6$, where the median ISE is 45% lower than the median ISE for the worst fit of $N = 2$.

[60] Theoretical lower bounds coming from the parameter vectors that make the estimated distributions as close as possible to the true distribution are ISE's of 0.0877, 0.0456, 0.0397, 0.0273, 0.0269 for $N = 0, 2, 4, 6, 8$.

Figure 5 provides a graphical illustration of the performance of our method. In each of the three panels we present the true distribution in bold and three estimated distributions corresponding to the simulations (using $N = 4$) that were at the 25^{th} percentile, the 50^{th} percentile, and the 75^{th} percentile in the MISE measure of goodness of fit. In the exponential and log-normal cases the estimated distributions seem to fit reasonably well for values of around the mean ($u = 1$) and to fit quite well for higher values of u. The estimated distributions are farther from the truth at low values of u. This should be expected – once we are considering a population of schools in which all schools will in practice have zero or one high-scoring student per year, a single year's data will not allow one to say whether all schools are identical or whether there is heterogeneity.

Also as expected, our method performs somewhat poorly for the uniform distribution with its bounded support. However, we are encouraged to note that, even for this difficult case, the estimated distribution does mostly spread out the mass over the correct $[0, 2]$ interval.

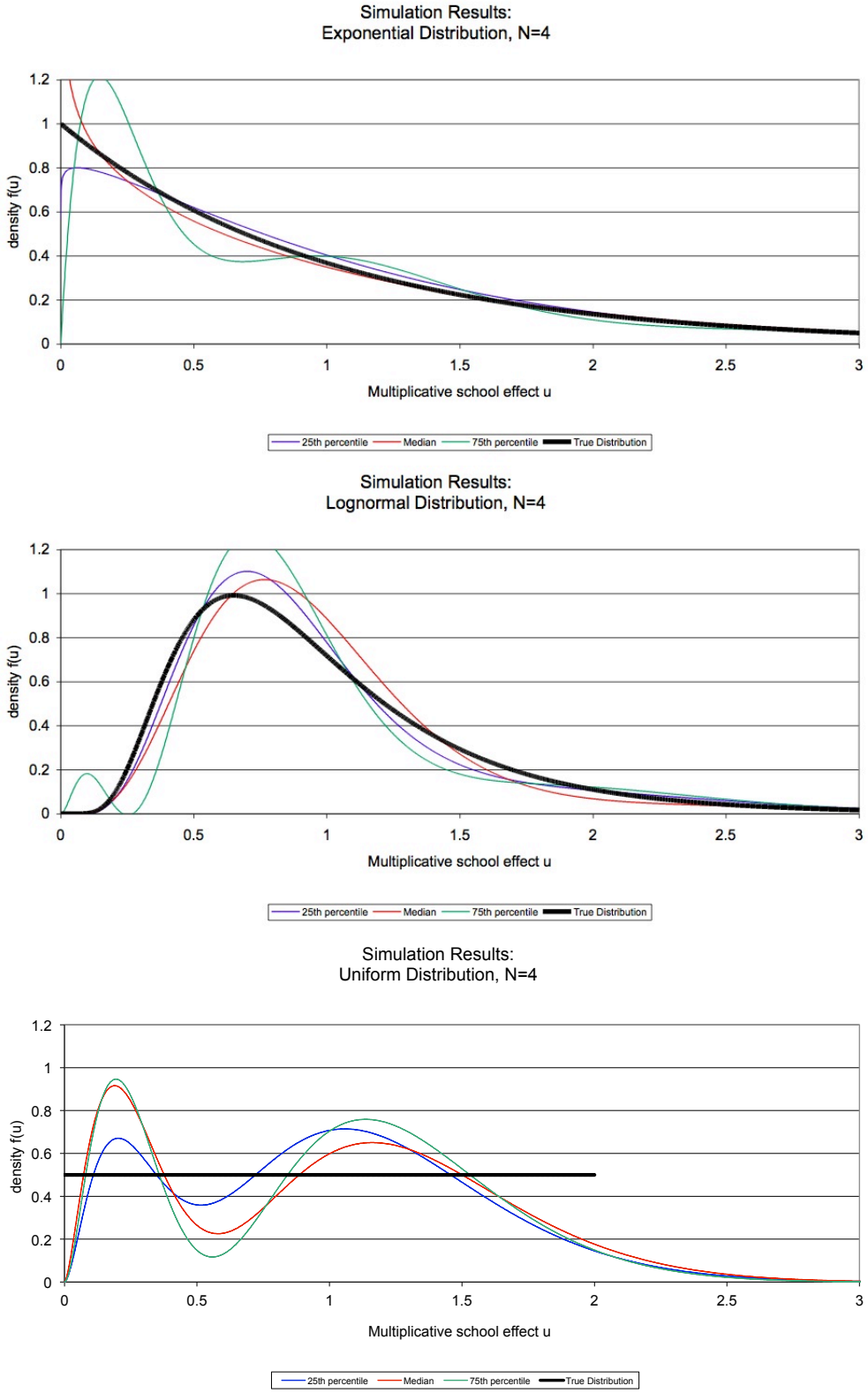

Figure 5: Actual vs. Estimated Distributions: 25^{th}, 50^{th}, and 75^{th} percentile fits in simulations