Ellison, Glenn; Swanson, Ashley

Working Paper

Heterogeneity in high math achievement across schools: Evidence from the American Mathematics Competition

CESifo Working Paper, No. 3903

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Ellison, Glenn; Swanson, Ashley (2012) : Heterogeneity in high math achievement across schools: Evidence from the American Mathematics Competition, CESifo Working Paper, No. 3903, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/61012

CESifo Working Papers
www.cesifo.org/wp

Heterogeneity in High Math Achievement Across Schools: Evidence from the American Mathematics Competition

Glenn Ellison
Ashley Swanson

CESIFO WORKING PAPER NO. 3903
CATEGORY 5: ECONOMICS OF EDUCATION
JULY 2012

CESifo *Working Paper No. 3903*

Heterogeneity in High Math Achievement Across Schools: Evidence from the American Mathematics Competition

Abstract

This paper explores differences in the frequency with which students from different schools reach high levels of math achievement. Data from the American Mathematics Competitions is used to produce counts of high-scoring students from more than two thousand public, coeducational, non-magnet, non-charter U.S. high schools. High-achieving students are found to be very far from evenly distributed. There are strong demographic predictors of high achievement. There are also large differences among seemingly similar schools. The unobserved heterogeneity across schools includes a thick tail of schools that produce many more high-achieving students than the average school. Gender-related differences and other breakdowns are also discussed.

JEL-Code: I210.

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

Glenn Ellison
Massachusetts Institute of Technology
Department of Economics (MIT)
USA - Cambridge, MA 02142
gellison@mit.edu

Ashley Swanson
Massachusetts Institute of Technology
Department of Economics (MIT)
USA - Cambridge, MA 02142
a.t.swan@gmail.com

July 2012
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 provided outstanding research assistance. Victor Chernozhukov provided important ideas and help with the methodology. We 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.

1 Introduction

There has been a great deal of recent interest in educational productivity. Understanding differences across schools may provide insights into improvements that educational reforms could bring. Where most of this literature focuses on differences in average achievement levels, we explore heterogeneity in the rates with which different schools produce high-achieving math students. We note that there are large differences related to student demographics and estimate the distribution of unobserved school effects after controlling for some of these. We find substantial heterogeneity among schools with similar demographics. This includes a thick upper tail of schools that produce many more high-achievers than would be expected given their demographics.

Roughly one can think of the "high achieving" math students we study as students who would score 800 on the math portion of the SAT reasoning test.[1] Several motivations can be given for studying the rate at which schools produce such students. First, high-achieving students make important contributions to scientific and technical fields, to human-capital intensive industries, etc.[2] Second, contrary to the impression one might get from the popular press, the poor performance of U.S. math education is not limited to leaving some students behind: international comparisons find that the U.S. trails most OECD nations in the production of high-achieving math students (but not high-achieving students in reading).[3] Third, policies may have very different effects on high-achieving and average students.[4]

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

[1] 800 is the highest possible score on the math SAT and is achieved by approximately 1 percent of SAT-takers.

[2] See Hoxby (2002) for a quick discussion of educational productivity as a source of U.S. comparative advantage. Krueger and Lindahl (2001) and Hanuschek and Woessmann (2008) provide surveys of the empirical literature on education and growth. Some 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, Google's Sergey Brin, who finished in the top 55 on the 1992 Putnam Exam, and Facebook's Mark Zuckerberg, who finished 13th on the Algebra test at the 2001 Harvard-MIT Math Tournament.

[3] 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 perentage of "Level 6" reading students is well above the OECD average (1.5% vs. 0.8%).

[4] See Neal and Schanzenbach (2010) and Dee and Jacob (2009) for recent empirical studies of such differences.

The test is explicitly designed to distinguish among students at very high achievement levels.[5] Our primary measures of high achievement will be the number of students in each school achieving AMC 12 scores above various cutoffs. We match AMC schools to schools surveyed by the National Center for Education Statistics (NCES) and to census data to obtain other covariates and conduct most of our analyses on the set of public, coed, non-magnet, non-charter U.S. high schools that administer the AMC 12 and could be matched to NCES data. Section 2 discusses the data in more detail.

The most basic question we address is whether there are substantial differences across schools in the production of high-achieving math students (after one controls for demographic differences). Going back to Coleman (1966), many papers have emphasized that differences in school-mean test scores are relatively "small" and much of the differences that do exist can be explained by demographic differences in the student populations.[6] In contrast, the literature on teacher value added tends to emphasize that there is substantial heterogeneity across teachers within a school.[7]

Our most basic finding is that there is a great deal of variation in the number of high-achieving math students produced by schools with similar demographics. We present several pieces of related evidence in Section 3. We begin with a simple informal comparison of high-achieving and matched comparison ZIP codes. Most of the section is then devoted to negative binomial regressions examining how the number of students with high AMC 12 scores is related to demographics and school characteristics. Demographics are found to be very strong predictors of high math achievement. But substantial heterogeneity remains after controlling for several variables. One noteworthy demographic finding is the income-achievement relationship: income is positively correlated with high achievement in the full national sample, but once we restrict to the relatively high quality schools which offer the AMC tests, there are fewer high-achieving math students in more affluent areas.

Just as the literature on heterogeneity in teacher quality has produced interesting graphs of the distribution of teacher value added, another objective of our paper is to provide estimates of the underlying *distribution* of unobserved heterogeneity across schools.[8]. For example, we estimate how many schools produce high-achieving students at less than one-

[5] See Ellison and Swanson (2010) for some evidence that the test is able to make meaningful distinctions.

[6] See Kane and Staiger (2002) for a discussion of differences across schools. Rothstein (2005) reports that four demographic variables reflecting racial composition and parental education and a control for participation rates account for 80% of the variation in school-average SAT scores.

[7] See, for example, McCaffrey et al. (2004), Aaronson, Barrow, and Sander (2007), Rivkin, Hanushek, and Kain (2005), Kane and Staiger (2008), and Chetty et al. (2011).

[8] See 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.

half of the expected rate, how many produce such students at more than twice the expected rate, and so on. Counts of high-achieving students are inherently small, so there will unavoidably be a great deal of randomness in the assessment of any individual school using a single year's data. Section 4 describes the methodology that we use to take out this random variation and thereby back out estimates of the underlying heterogeneity. The method involves a series expansion similar to that of Gurmu, Rilstone, and Stern (1999), but relying on a different characterization of the likelihoods. Appendix I contains more detail on the estimation along with Monte Carlo estimates illustrating the performance of the estimator.

Section 5 presents estimates of the distribution of unobserved heterogeneity across schools. Our main qualitative findings are that many schools produce high-achieving students at much less than the average rate and that there is a strikingly thick tail of extremely successful schools. For example, we estimate that about 38% of schools produce high-scoring students at less than one-half of the average rate for schools with similar characteristics, and that 2% of schools produce high-scoring students at more than five times the average rate. We also provide estimates of the heterogeneity of school effects relevant to subpopulations, including the likelihood of producing high-achieving female students. Here, the estimates suggest that many schools are extremely unlikely to produce high-achieving girls.

Part of the outcome heterogeneity described in Section 5 is due to the self-selection of high-ability students into certain public schools. Section 6 examines another data source that may provide some insight on this effect: it examines differences in the rates at which different schools produce students with extremely high SAT scores. We find that the estimated distribution of unobserved heterogeneity derived from this measure of high achievement does not have the thick tail we see in the estimates derived from AMC data. If self-selection of high-ability students is equally or more important in the production of students with very high SAT scores, this would suggest that self-selection is not responsible for the thick upper tail in the AMC data.

Our work is related to a number of literatures. One is the literature on quality differences across schools affecting average achievement. This includes papers examining how inputs affect achievement (Coleman (1966), Hanushek (1986), Card and Krueger (1992), etc.) and papers that focus on differences in productivity related to competition (Hoxby (2000)), vouchers (Hoxby (2000, 2002)) and charter schools (Angrist et al. (2002), Dobbie and Fryer (2009), Hoxby et al. (2009)). Many of these papers control for selection effects and estimate causal effects relevant to school reform debates.

There are also a number of papers that discuss high-achieving students. Hanushek, Peterson, and Woessmann (2011) note that PISA data indicate that the U.S. trails most OECD countries in the fraction of students reaching high levels of math achievement. Using a concordance between PISA and NAEP scores, they note that while there are large differences across U.S. states, even the best performing U.S. states (and high-performing demographic groups) have a lower percentage of high achievers than do many countries. A number of papers have noted that proficiency-based reforms create an incentive for schools to focus on students near the cutoff and that resource diversion could harm high-achieving (and very low-achieving) students and examine such effects empirically.[9] Bui, Craig, and Imberman (2011), Abdulkadiroglu, Angrist, and Pathak (2011), and Dobbie and Fryer (2011) examine the effects of gifted programs and magnet schools on (marginal) high ability students using regression discontinuity designs. Andreescu *et al.* (2008), focuses on a much higher percentile (roughly the 99.9999th), and has a message related to ours in noting that the highest achieving students are a highly nonrepresentative sample of the U.S. population ethnically and in the schools they attend.

The methodological part of our paper contributes to the econometric literature on count data models. This literature contains many alternatives to the Poisson model that can better fit datasets in which conditional means and variances are not equal. For example, the classic negative binomial model does this by allowing for gamma-distributed unobserved heterogeneity.[10] Two approaches to semiparametric estimation have been developed for applications where (as in our case) the distribution of underlying heterogeneity is an object of interest.[11] Brännäs and Rosenqvist (1994) develop an estimator along the lines of Simar (1976) and Heckman and Singer (1984) which involves modeling the distribution of unobserved heterogeneity as a finite number of mass points. Gurmu, Rilstone and Stern (1999) develop a series estimator which involves representing the unobserved heterogeneity as the product of a gamma-distributed density and function which is flexibly represented via an orthogonal polynomial expansion. Our primary motivation for not simply adopting one of these estimators is that previous Monte Carlo studies have suggested that they may not

[9] See Krieg (2008), Neal and Schanzenbach (2010), Dee and Jacob (2009). These papers assess performance using state proficiency tests and/or the NAEP and focus on students at the 90th percentile as their high-achieving population.

[10] Several other models allow for unobserved heterogeneity of other forms, and there are also approaches that focus directly on flexibly estimating discrete distributions without an underlying Poisson model. See Cameron and Trivedi (1998), Guo and Trivedi (2002), and Winkelmann (2008) for overviews.

[11] There are also semiparametric estimators in other branches of the count-data literature, e.g. Cameron and Johansson (1997).

work well in our application.[12] Our approach is similar to that of Gurmu *et al.* (1999) in that we also specify the distribution of underlying heterogeneity as a product of a gamma-like density and a flexible orthogonal polynomial term using Laguerre polynomials. The main difference is that we exploit other properties of the Laguerre polynomials to derive a different expression for the probability of each outcome.[13] Our expression does not involve the moment generating function and Monte Carlo estimates indicate that it can work reasonably well even with a fat-tailed distribution.

2 Data

The primary subject of our analysis is a database of scores on the Mathematical Association of America's AMC 12 contest from 2007.[14] The AMC 12 is a 25-question, multiple choice test that is explicitly designed to identify and distinguish among students at very high performance levels. It is taken by a self-selected sample of mostly high-achieving students, but the average score among U.S. students is just 66.3 of 150.[15] We will focus primarily on very high-achieving students who scored above 100 on the exam. As discussed in Ellison and Swanson (2009), scoring 100 on the AMC 12 can be thought of as comparably difficult to scoring 800 on the math SAT. The primary advantage of the AMC 12 is that it is more reliable in this range and can also draw consistent distinctions at higher percentiles. Approximately 5% of U.S. participants scored in excess of 100 on the AMC 12 (which places them approximately in the 99th percentile of the SAT-taking population) and approximately 0.5% scored in excess of 120.

Our raw data are at the individual level, but only minimal information about each student is available – age, grade, gender, home ZIP – so we will mostly aggregate the

[12] Heckman and Singer (1984) report that their method is better at estimating structural parameters than at estimating distributions of heterogeneity; and Gurmu *et al.* (1999) find that they do not estimate the distribution of heterogeneity well when the underlying distribution is log-normal (which may be relevant to our application because our data suggest fat tails). Their estimator does appear to provide reliable estimates of the systematic relationships for a broad range of distributions. Their Monte Carlo result on the difficulty in recovering a log-normal density using their estimator is not unexpected: their estimator can be seen as flexibly estimating the moment generating function of the distribution of heterogeneity. The log-normal distribution is sufficiently fat-tailed so that the moment generating function does not exist.

[13] Our approach also has several weaknesses 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.

[14] The AMC 12 is open to all students in grades 12 and below. Approximately 88% of the participants are in 11th or 12th grade.

[15] Students receive 6 points for each correct answer and 1.5 points for each answer left blank so scores range from 0 to 150.

data to the level of the high school (or home ZIP code) and treat schools as the unit of observation.[16] The primary school level variables we analyze are counts of the number of students in the school scoring at least 100 or 120.

We combine the AMC data with demographic information on participants' schools and areas of residence. ZIP code level demographics were taken from the U.S. Census. School level variables are constructed by matching to the National Center for Education Statistics (NCES) database for 2005-2006. The NCES collected private school data from schools that responded to the Private School Universe Survey (PSS). Data on public schools are from the NCES Common Core of Data, which is collected annually from state education agencies. School name, city, and state data were linked to the AMC data using the CEEB code search program provided on the College Board's website. CEEB codes for schools participating in the AMC 12 were then matched to NCES data by school name, city, and state. Of the 3,730 schools with numerical CEEB codes in the AMC data, 3,105 were matched to schools in the NCES data.[17] Among the variables we will use in our analyis are the number of students enrolled in each school, the percent of students belonging to various racial and ethnic groups, the percent of students qualifying for free lunch and the school's Title I status. We also use information from the NCES to restrict the sample to public, co-ed, non-magnet, non-charter U.S. high schools in most analyses.

3 Differences Across Schools: Magnitudes and Sources of Heterogeneity

We begin this section by noting a basic fact: there is tremendous variation in the number of high-achieving students who are coming out of different schools and home ZIP codes. We then examine the relationship between the number of high-scoring students in a school and observable demographic characteristics, noting both that there are a number of strong patterns and that there is a great deal of residual heterogeneity among seemingly similar schools.

[16] The AMC 12 contest is offered twice each year. The exams are different and students are allowed to take both, which about 2% of the participating students do. We matched such records as well as we could and included only scores from the later test date for dual takers.

[17] 311 of the remaining 625 schools do not appear in the NCES data because they are not in the U.S., and a further 158 could not be matched because they did not have official CEEB identifiers and were thus not linked to school data of any kind. It was not possible to match the remaining 156 schools by the identifiers reported by the College Board or NCES. Some of the remaining schools with valid CEEB codes may not appear in the NCES survey data because private schools are not required to fill out the PSS.

3.1 A first look at differences across ZIP codes

A simple way to illustrate the heterogeneity in student outcomes in the AMC data is to look at counts of the number of students residing in each ZIP code who scored least 100 on the AMC 12. There are about 32,000 residential ZIP codes in the U.S. and about 5,000 U.S. students scored over 100 on the 2007 AMC 12. If all students were equally likely to score 100 on the AMC 12, then (taking into account that some ZIP codes are more populous than others) we'd expect about 12 percent of ZIP codes to have at least one AMC high scorer: about 10% would have one high scorer, another 2 percent would have two high scorers and less than one percent of ZIP codes would have more. Extreme concentrations would be very rare: there is less than a 0.0001 chance that even a single ZIP code would have ten or more high scorers. Reality looks very different from this. Only 6% have any high scorers. And 58 ZIP codes have ten or more.

The raw differences noted in the previous paragraph reflect various factors: socioeconomic differences in the student populations; differences in school quality; and differences in whether interested students take the AMC 12.[18] Table 1 presents some additional data on the most successful ZIP codes. One of our main findings will be that there's a thick upper tail of extremely successful schools. The first ten rows of Table 1 illustrate this by listing the ten ZIP codes in which the highest number of AMC 12 high scorers reside. Each of these ZIP codes had at least twenty students scoring at least 100 on the AMC 12, which means they are more than 20 standard deviations above average. Although they are more populous than average, they are still all at least 4 standard deviations above average in the number of high scorers per capita.

The table also reports a few demographics. In addition to being more populous than the average ZIP code, the highest-achieving ZIP codes tend to have above-average incomes, many highly educated adults, and large Asian-American populations. The second-most successful ZIP code (Exeter, NH) is the location of an elite boarding school, but most others are just parts of some major metropolitan area in which most students attend a highly-regarded public school. Our regression analyses will focus on standard public schools.

The bottom half of Table 1 illustrates both that demographics have predictive power and another finding: a great deal of variation remains even when we compare areas/schools with similar demographics. In these rows we present some data on a comparison set of ZIP codes which are similar to those in the top half of the table. For each ZIP code in the

[18] Although the vast majority of students take the AMC 12 in their own high school, interested students may also be able to take the test at a nearby college or at a number of other locations.

top half we chose the ZIP code sharing the same first two digits which had the smallest weighted sum-of-absolute differences in the listed characteristics other than population.[19] The substantial number of high AMC scorers in the comparison ZIP codes suggests that there are strong demographic predictors of high AMC scores: the comparison ZIP codes average of 3.3 high scorers whereas the average across all ZIPs nationwide is just 0.2. But the comparison ZIP codes are still quite far behind the ZIP codes in the top half of the table: an illustration of our observation that there is a lot of unexplained heterogeneity in the upper tail.

3.2 Demographic patterns in high math achievement

In this section we aggregate our data to the level of the school or ZIP code and look at the relationship between the number of high-achieving students and school/ZIP code demographics.

The first column of Table 2 presents coefficient estimates (with t-statistics in parentheses) from a simple negative binomial regression with the number of students in the school scoring at least 100 on the AMC 12 as the dependent variable. A number of the strong effects match one's intuition. Parental education is very important: a one percentage point increase in the fraction of adults in the ZIP code with bachelors' and graduate degrees increases the expected number of AMC high-scorers by 3.1 and 5.9 percent, respectively. The school's ethnic makeup also matters in that a one percentage point increase in the Asian-American population of a school increases the expected number of AMC high scorers by over two percent. (Percent white is omitted from the regression.) There are also fewer high scorers in schools with more students qualifying for the free lunch program.

An effect that may not match intuition is the income effect: after controlling for the ethnic makeup of the school and parental education we find that there are fewer high scorers in higher-income areas. It should be kept in mind that the negative income effect is occurring after we restrict to schools offering the AMC and include the fraction of students qualifying for free lunch as a covariate. One thought on why this might occur, other than from a selection effect, is that social norms or the college application process may lead students in wealthier schools to spend more time on other activities, such as athletic and arts programs, debate teams, model UN, etc., rather than developing their math skills to a very high level.

The second column of the table looks at students reaching an even higher achievement

[19] The weights are taken from a regression model predicting the number of high scorers.

City	ZIP Code	# AMC12 Over 100	# per 10K pop.	Pop.	Med. Income	% Grad.	% Asian
ZIP Codes with the Largest Number of AMC 12 High Scorers							
San Jose, CA	95129	43	11.4	37,674	79,489	24.8	41.3
Exeter, NH	03833	28	14.6	19,129	51,858	16.3	1.0
Sugar Land, TX	77479	27	4.8	55,682	96,118	21.3	22.9
Saratoga, CA	95070	27	8.8	30,590	138,206	33.3	28.3
Fremont, CA	94539	25	5.3	46,997	101,977	26.5	49.5
Naperville, IL	60540	25	5.9	42,100	87,514	26.4	8.7
Naperville, IL	60565	21	5.2	40,503	97,807	22.9	9.7
Gaithersburg, MD	20878	21	3.8	55,186	84,330	29.6	18.7
Bayside, NY	11364	20	5.8	34,575	54,031	15.1	32.9
McLean, VA	22101	20	7.0	28,550	125,105	45.0	10.9
Matched Comparison ZIP Codes							
Santa Clara, CA	95054	0	0.0	12,860	85,124	18.1	46.1
Spofford, NH	03462	0	0.0	1,729	50,885	16.3	0.0
Missouri City, TX	77459	5	1.5	32,774	84,901	17.9	14.0
San Jose, CA	95120	5	1.3	37,175	120,117	26.2	22.9
Fremont, CA	94555	2	0.6	33,811	84,442	18.6	53.8
Northbrook, IL	60062	5	1.2	40,175	89,164	26.7	10.4
Hinsdale, IL	60521	3	0.8	37,489	91,727	25.2	7.7
Rockville, MD	20850	5	1.5	33,277	74,655	31.1	17.3
Glen Oaks, NY	11004	1	0.7	14,760	55,156	14.7	29.7
Vienna, VA	22182	7	3.1	22,758	120,075	36.4	13.0
All U.S. ZIPs	Mean	0.2	0.1	8,901	39,394	6.3	1.4
	(Std. Dev.)	(0.9)	(0.9)	(13,105)	(16,427)	(6.6)	(4.1)

Notes and sources: Characteristics of the ten ZIP codes with the highest number of AMC 12 high scorers and ten comparison ZIP codes. AMC 12 high scorers are students scoring more than 100 on the AMC 12. Each comparison ZIP code identified based on having minimum weighted sum-of-absolute differences in listed characteristics other than population among other ZIP codes with same two-digit ZIP code prefix as top-ten ZIP code. Data from 2000 United States census.

Table 1: ZIP codes with the most AMC 12 high scorers and comparison ZIP codes

Variable	AMC 12 High Scorers				High SAT Scorers	
log(Num. of Students)	1.01	1.32		1.05	1.17	1.18
	(13.33)	(6.20)		(12.22)	(11.51)	(10.78)
Adult frac. BA	3.14	5.36	4.78	3.18	3.66	3.72
	(4.71)	(3.46)	(10.62)	(4.72)	(4.69)	(4.48)
Adult frac. Grad	5.94	7.27	6.97	6.29	6.53	6.61
	(11.13)	(6.67)	(17.46)	(12.41)	(11.82)	(11.60)
log(ZIP median income)	-0.94	-1.36	0.73	-0.90	-1.17	-1.17
	(6.88)	(4.39)	(8.29)	(6.72)	(7.75)	(7.74)
ZIP frac. urban	0.52	0.43	0.32	0.55	0.31	0.31
	(2.48)	(0.68)	(2.42)	(2.34)	(1.26)	(1.17)
School/ZIP frac. Asian	2.08	2.44	3.70	2.00	1.19	1.15
	(7.48)	(4.17)	(12.61)	(7.51)	(3.91)	(3.80)
School/ZIP frac. Black	-0.05	0.15	-1.10	-0.39	0.09	0.06
	(0.15)	(0.17)	(5.31)	(1.08)	(0.21)	(0.13)
School/ZIP frac. Hisp.	-0.38	-1.81	-1.59	-0.77	-1.57	-1.73
	(1.16)	(1.81)	(7.12)	(2.04)	(3.11)	(3.18)
School frac. female	0.01	2.92		0.30	2.28	2.66
	(0.00)	(0.83)		(0.20)	(1.41)	(1.56)
Title 1 school	-0.04	0.12		-0.11	-0.12	-0.14
	(0.48)	(0.56)		(1.15)	(1.08)	(1.15)
School free lunch frac.	-2.39	-2.92		-2.22	-2.75	-2.71
	(5.46)	(2.30)		(4.62)	(4.59)	(4.37)
log(Population)			1.10			
			(30.37)			
log(Award ratio)					0.83	0.84
					(15.21)	(14.44)
Constant	1.15	-0.92	-22.16	0.32	6.47	6.32
	(0.70)	(0.23)	(22.34)	(0.96)	(3.53)	(3.38)
Threshold	100	120	100	100		
Unit of obs.	School	School	ZIP	School	School	School
Pseudo R^2	0.1492	0.1627	0.3463	–	0.1985	–
Est. Method	NB	NB	NB	semi-P	NB	semi-P
$\hat{\alpha}$	1.02	2.77	1.42	–	0.48	–
(Std. Error)	(0.07)	(0.48)	(0.07)	–	(0.08)	–
# of obs.	2,165	2,165	31,889	2,165	2,165	2,165
# of high scorers	3,000	315	4,929	3,000	1,044	1,044

Notes and sources: Results of negative binomial regression and semi-parametric model estimation. Outcomes are counts of high achievers (students scoring more than 100 on the AMC 12, students scoring more than 120 on the AMC 12, and students achieving PSP candidate status on the SAT) in each school or ZIP code. School demographics are from the NCES Common Core of Data for 2005-6; ZIP code demographics are from the 2000 U.S. census. Award ratio is the ratio of the number of PSP candidates to the number of public and private high school graduates in each state.

Table 2: High Math Achievers vs. School Characteristics

threshold: scoring 120 on the AMC 12. (They can be thought of as above the 99.9th percentile among college-bound students).[20] Given that most students who would score 120 on the AMC are probably taking the test, the results here should be less affected by differences in how widely administered the AMC is at each school. The effects noted in the previous paragraph remain present and generally increase in magnitude, suggesting that the earlier results were not due to this type of selection.

Our school-level estimates reflect a highly nonrepresentative sample of the U.S. population: we only include schools offering the AMC and these are disproportionately high-achieving schools located in relatively wealthy areas. To give a sense of where AMC high scorers are coming from relative to the whole U.S. population, the third column of Table 2 goes back to the ZIP code as the unit of observation and presents estimates from a negative binomial regression with the number of resident students scoring at least 100 on the AMC 12 as the dependent variable.[21] Coefficients here should be thought of as reflecting both differences in math achievement and differences in access to the AMC. The effects of greater education in the adult population become larger in these results, as we might expect given that the availability of the AMC may be correlated with parental education. The coefficient on income becomes positive and highly significant. A couple of factors may be responsible for the change. One is that access to the AMC may be increasing in income. Another is that there may be a nonmonotonic relationship between income and investments in reaching high levels of math achievement: it may increase through much of the income distribution (all of which is relevant for this regression) and then turn down at the very top (which is more relevant once we restrict to AMC-offering schools).

3.3 Magnitudes of unobserved heterogeneity

The negative binomial regression model also provides a simple estimator of the degree of unobserved heterogeneity that remains across schools after controlling for a given set of variables. We present some such estimates here.

Recall that the negative binomial regression model assumes that Y_i is a Poisson random variable with mean $e^{X_i\beta}u_i$, where u_i is an independent mean one gamma-distributed random variable reflecting underlying heterogeneity not captured by the observed X_i. The negative binomial model estimates both the coefficients β on the X's and an extra parameter α which is the variance of the u_i. Estimates of this parameter from each of the negative

[20] Given our restriction to public, non-magnet, non-charter, coeducational schools, there are 315 students in our sample at this threshold.

[21] Here, the ethnic composition variables are for the student's home ZIP code rather than the school.

binomial models discussed in the previous section are presented in Table 2.

To provide more of a feel for the magnitude of the unobserved heterogeneity relative to the heterogeneity captured by our demographic controls Table 3 reports the estimated standard deviation of u (the square root of $\hat{\alpha}$) from negative binomial regressions of the number of high-scoring AMC students in a school/ZIP code on different sets of controls.

The first row presents results with almost no controls: the number of high scorers in a ZIP code is the dependent variable and the only control is the ZIP code's population. The standard deviation of u is 2.59 for the count of students scoring over 100, and 3.92 for the count of students scoring over 120, indicating substantial unobserved heterogeneity in the number of high scorers beyond what can be explained by population variation alone. The estimated $SD(u)$ is much greater for the higher score threshold, indicating that unobserved factors are more important at higher performance levels. The second row illustrates the effect of introducing the same demographic controls used in the ZIP code regressions displayed in Table 2. Introducing controls reduces the unobserved heterogeneity by more than half for each score threshold. The unobserved heterogeneity remains large in a practical sense and remains highly statistically significant.[22] One might imagine that the unobserved heterogeneity is entirely due to differences between areas that offer the AMC and areas that do not. To examine this hypothesis, the third row restricts the sample to ZIP codes that had at least one student take the AMC 12; this restriction reduces the estimated $SD(u)$ only slightly. The final two rows use schools as the unit of observation and restrict attention to public, non-magnet, non-charter, coeducational U.S. high schools in which the AMC 12 was offered. These estimates are in the same range as those observed for the ZIP-level analyses and exhibit similar patterns in that the measure of dispersion is higher for the higher thresholds and lower once we introduce controls. Even the numbers in the bottom row, however, are again very large in a practical sense. For example, the standard deviation of the unobserved heterogeneity would be 1 if 50% of all schools gave students no chance of achieving a high score on the AMC 12 and the other half gave twice the average chance, and it would be two if 80% of schools gave students zero chance of reaching high-achievement levels and 20% gave students five times the average chance. For the true values to be 1.01 or 1.66 there must be many schools which give students very little chance of reaching high achievement levels and other schools that give a much better than average chance.

An overall impression from these data is that there are several factors that are very strong predictors of whether a school/ZIP code will have many high-achieving math stu-

[22] χ^2 tests of the null that $SD(u) = 0$ are rejected at the 0.001 level in each case.

dents, but that there are also substantial differences across seemingly similar schools.

AMC 12 > 100				AMC 12 ≥ 120			
Student Groups	Sample	Controls	Est. SD(u)	Student Groups	Sample	Controls	Est. SD(u)
ZIP Codes	All	Pop. Only	2.59	ZIP Codes	All	Pop. Only	3.92
ZIP Codes	All	Yes	1.19	ZIP Codes	All	Yes	1.65
ZIP Codes	AMC Taking	Yes	1.12	ZIP Codes	AMC Taking	Yes	1.61
Schools	Public	Enrollment	1.46	Schools	Public	Enrollment	2.75
Schools	Public	Yes	1.01	Schools	Public	Yes	1.66

Notes and sources: Estimated standard deviation of performance heterogeneity from negative binomial regressions with and without demographic controls. School demographics are from the NCES Common Core of Data for 2005-6; ZIP code demographics are from the 2000 U.S. census.

Table 3: Magnitude of unexplained variation u

4 Methodology for Estimating the Distribution of Unobserved Heterogeneity

In this section we describe how we estimate the distribution of unobserved heterogeneity across schools and provide some simulation results.

4.1 Estimation method

Suppose the count variable 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 *et al.* (1999) in that we use a series expansion and exploit known properties of the orthogonal polynomials involved to facilitate maximum likelihood estimation. 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 $\int_{x=0}^{\infty} x^\alpha e^{-x} |g(x)|^2 dx < \infty$, 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, $L_j^{(\alpha)}(x) \equiv \sum_{i=0}^{j}(-1)^i\binom{j+\alpha}{j-i}\frac{x^i}{i!}$.[23] 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 1 *Consider the model described above. Then,*

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

The derivation of the formula exploits several properties of the Laguerre polynomials. Details are given in the 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.[24] For the estimated $f(u)$ to be a valid density, 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
- $u^\alpha e^{-u}\sum_{j=0}^{N} g_j L_j^{(\alpha)}(u) \geq 0$ 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. The second constraint is not as easy to express as a parameter restriction. We will not impose it as a constraint, but do add a penalty function (described in the Appendix) to the likelihood when the density is not everywhere positive, which in practice has the effect of making the estimated densities at most slightly negative.[25]

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

[23] The coefficients g_j are given by

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

[24] 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 $\|g-g^N\| \leq \sum_{j=N+1}^{\infty}\binom{j+\alpha}{j}g_j^2$.

[25] Our specification of the model also includes the scaling assumption that $E(u_i|z) = 1$. This is necessary for identification when the set of explanatory variables z contains a constant and the distribution of u is to be estimated semiparametrically. This condition can also be easily imposed as a parameter restriction: it is satisfied if and only if $g_1 = \alpha/\Gamma(\alpha+2)$. When only a finite number N of terms are included in the series expansion, however, imposing this constraint is not necessary for identification, and the restriction is not imposed in the estimates reported in this paper. Instead, we estimate the model without the restriction and just renormalize the estimated distributions of u to have mean 1 by dividing by their expectations when graphing them.

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 $\sum_{j=0}^{N} g_j L_j^{(\alpha)}(x)$. 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 ($\alpha \in (-1, 0)$).[26]

4.2 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.[27] 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 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.

5 Distributions of Unobserved Heterogeneity Across Schools

We noted in section 3 that there are substantial performance differences across schools with similar demographics. In this section, we explore the distribution of unobserved heterogeneity in more detail. The estimates will quantify our earlier informal remarks that there is a thick upper tail of schools that are much more successful than the average school, and provides a number of other insights.

[26] The pure Poisson model with no unobserved heterogeneity is obtained as a special case as $\alpha \to \infty$.
[27] See Gurmu *et al.* (1999) p. 141.

5.1 Differences across seemingly similar schools

We begin by estimating a model of the production of AMC 12 high-scorers similar to the benchmark negative binomial regressions of Section 3, but employing the methodology described in Section 4. Specifically, we estimate the model using the number of students in each school scoring at least 100 on the 2007 AMC 12 as the dependent variable and use the same demographics as in Table 2 as control variables. Again, the sample is the set of coed public, non-magnet, non-charter schools offering the AMC 12 that we were able to match to the NCES data.

Our primary interest in this section will be on the distribution of unobserved school effects. 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, a value of $u = 0.5$ corresponds to a school that produces high-scorers at half of this rate, etc. The curve is like a histogram giving the relative frequency of the values of u in the population of schools. The substantial differences between schools with similar demographics are clearly visible in the figure: it looks nothing like a distribution that is highly concentrated around $u = 1$. Instead, it is a spread-out distribution that is skewed to the right. There are a large number of schools that produce AMC 12 high scorers at well below the average rate. For example, about 38% of schools are estimated to produce high scorers at less than one-half of the average rate. At the other extreme, there is a tail of highly successful schools. For example, about 9% of schools are estimated to produce high scorers at more than double the average rate. The dashed lines in the figure are 95% confidence bands for the estimated density.[28] They indicate that the estimates are quite precise throughout most of the range, and then become much less precise in the lower tail.

While the density function looks roughly like an exponential in the range that is graphed, there is a notable departure in the right tail – the upper tail of the estimated distribution is much thicker than that of an exponential distribution (or normal distribution). The bottom panel of Figure 1 illustrates this by graphing in bold the CDF of the estimated distribution for u's ranging from 3 to 10. The estimates indicate that there are a substantial number of schools which produce high-achieving math students at five to ten times the average rate for a school with their demographics. The dashed lines again give a 95% confidence

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

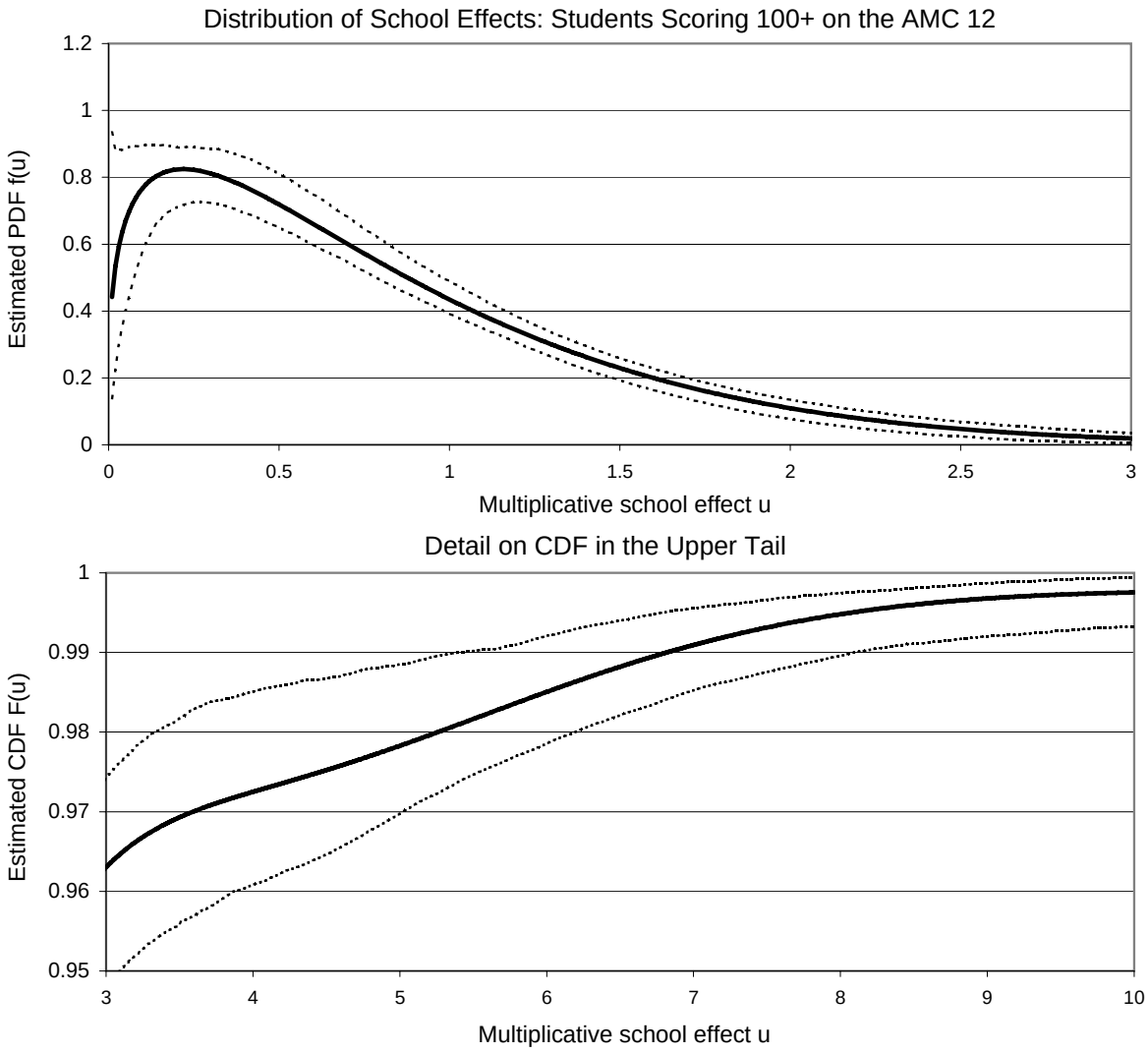

Figure 1: Estimated distribution of school effects: AMC 12 high scorers

interval. They indicate that the thick tail is a statistically significant phenomenon.

Our semiparametric estimation procedure also yields estimated coefficients on the demographic variables, which are presented in the fourth column of Table 2. The estimates are very similar to the estimates from the negative binomial regression as shown in the first column of the same table. The main implication to take away is that our earlier negative binomial regression results appear to be robust to modeling the unobserved heterogeneity more flexibly.

5.2 Female students

Given the underrepresentation of female students among high math achievers it seems natural to explore differences in how often schools produce high-achieving female students. In this section, we present estimates similar to those in the previous section, but focusing on how often schools produce high-achieving girls. Our main observation is that unobserved school quality appears to be even more important for girls than it is for boys: there is an upper tail of schools that produce high-scoring girls at more than ten times the average rate.

One way to quantify the increased importance of unobserved school effects for girls is to estimate a negative binomial regression similar to those used in Section 3 using the number of high-scoring female students per school as the dependent variable. The estimate $\hat{\alpha}$ from such a regression (available on request) is 1.88 (s.e. 0.26) whereas it was 1.02 (s.e. 0.07) when we examined high-scoring students of either gender.

Figure 2 presents graphs of the estimated distribution of the unobserved differences u across schools estimated using data on the number of female students scoring at least 100 on the AMC 12. The top panel shows the PDF for u's in $[0, 3]$ and the bottom panel shows the estimated CDF for the upper tail of schools. Again, the bold lines are the estimated PDFs and CDFs of the school effects relevant to girls and the thinner dashed lines are 95% pointwise confidence bands.

The estimated distribution is qualitatively similar to what we reported in our earlier analysis of how often schools produced high-scoring male or female students in a couple respects. First, there are a large number of schools where girls are relatively unlikely to reach high levels of math achievement. For example, about 28% of schools are estimated to produce high-scoring girls at less than one-fourth of the average rate. Second, there is a small but very thick upper tail of schools where girls are many times more likely to succeed than are girls in an average school with comparable demographics.

In Ellison and Swanson (2010) we noted that there is a large gender gap among high-achieving students in the AMC data: there is roughly a 4:1 male-female ratio among student scoring at least 100 on the AMC 12; and the average number of girls per school reaching this level is just 0.26. Hence, the estimate that there are a substantial number of schools that produce high-scoring girls at less than one-fourth of the average rate is very discouraging: it implies that these schools produce high-scoring girls at a rate of well less than one per decade. But the presence of the right tail is encouraging. The 99^{th} percentile school is estimated to produce high-scoring girls at more than ten times the rate of the average school with comparable demographics. This result implies that there are a number of schools that produce high-scoring girls at a rate that surpasses the rate at which the average school produces high-scoring boys.

The estimated distributions suggest that there are more schools in the lower and extreme upper tails for girls. But the imprecision of the estimates makes it difficult to make statistically significant statements about where exactly in the distribution the extra variance is coming from. A number of potential explanations could be given for why there might be more schools in the lower tail when we examine 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. Another plausible story might be that girls are relatively disadvantaged when a school's "honors" math classes are not taught at a very high level (perhaps because girls are less liable to complain or take supplementary online classes).

5.3 School effects at higher achievement levels

The AMC data also make it possible to study the distribution of students at even higher math achievement levels. Here, we present an estimated PDF of school effects estimated using the number of students scoring at least 120 on the AMC 12. We have two motivations. First, it seems natural to take advantage of a relatively rare opportunity to examine where $99.9+^{th}$ percentile students are coming from. Second, looking at such a high percentile will help to purge the results of one selection effect. It is unlikely that much of the differences across schools could be due to differences in the degree to which schools encourage their high-achieving students to take the AMC 12 because it is unlikely that students would reach such a high level of mastery if they were not planning to participate in contests like the AMC. Recall that the negative binomial regressions presented in Table 2 indicated that there was more unobserved heterogeneity across schools when we examined students scoring

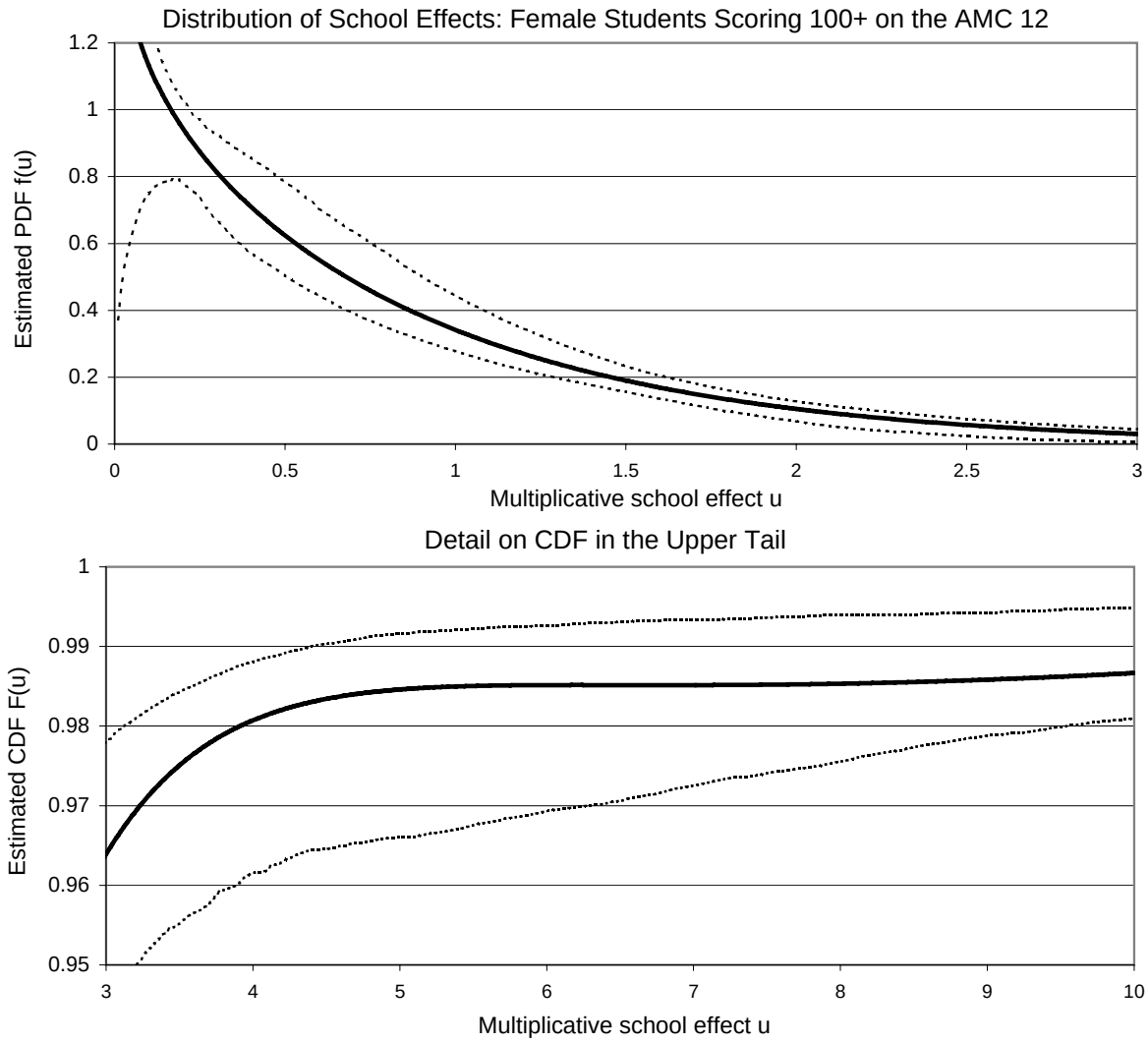

Figure 2: Distribution of school effects: female AMC 12 high scorers

at least 120 on the AMC 12.

The top panel of Figure 3 presents the estimated PDF for u's between 0 and 3, and the bottom panel shows the upper tail of the CDF. Again the estimates are in bold with 95% pointwise confidence bands around them. The estimated distribution of school effects once again has a very thick upper tail. This supports the hypothesis that the upper tail heterogeneity is not just an artifact of selection into taking the AMC test. Comparing the estimated density to that estimated earlier from data on students scoring at least 100, we find two main differences: more schools are now estimated to be very far below average; and more schools are now estimated to produce high-scoring students at three to five times the average rate. Again, however, it is difficult to make statistically significant pointwise comparisons. Certainly, however, the combination of the low mean number of students scoring 120 and the large mass in the lower tail means that there are a large number of schools in which it is extremely unlikely that students will reach the very high levels of math achievement considered here.

5.4 Heterogeneity in the full U.S.: ZIP code breakdowns

In this section we reestimate our model using ZIP codes rather than schools as the unit of observation. In this way, we are able to provide estimates of the unexplained heterogeneity in high math achievement throughout the country (whereas previous estimates compare schools to other schools offering the AMC 12 test).

The estimated heterogeneities in this section will reflect both differences in the number of high-achieving students in a ZIP code and differences in AMC participation rates among such students. The participation rate differences should be smaller for students at very high percentiles of achievement. Accordingly, we focus in this section on students scoring at least 120 on the AMC 12.

Figure 4 presents estimated distributions of ZIP code effects obtained by estimating the model using a count of the number of students scoring at at least 120 as the dependent variables and using ZIP code characteristics as control variables.[29] The top panel gives the PDF for values of u between 0 and 3 and the bottom panel gives a magnified view of the upper tail of the CDF. The distribution is similar to many others we've seen. There is a very thick left tail of ZIP codes in which students are estimated to have a much lower than

[29]The ZIP code characteristics are those in Table 2: log of population, percent of population with a bachelor's degree, percent of population with a graduate degree, percent urban, percent of population which is Asian, black, Hispanic, and female, and log of median income. The model is estimated on the full sample of 31,889 residential ZIP codes for which characteristics data were available in the Census.

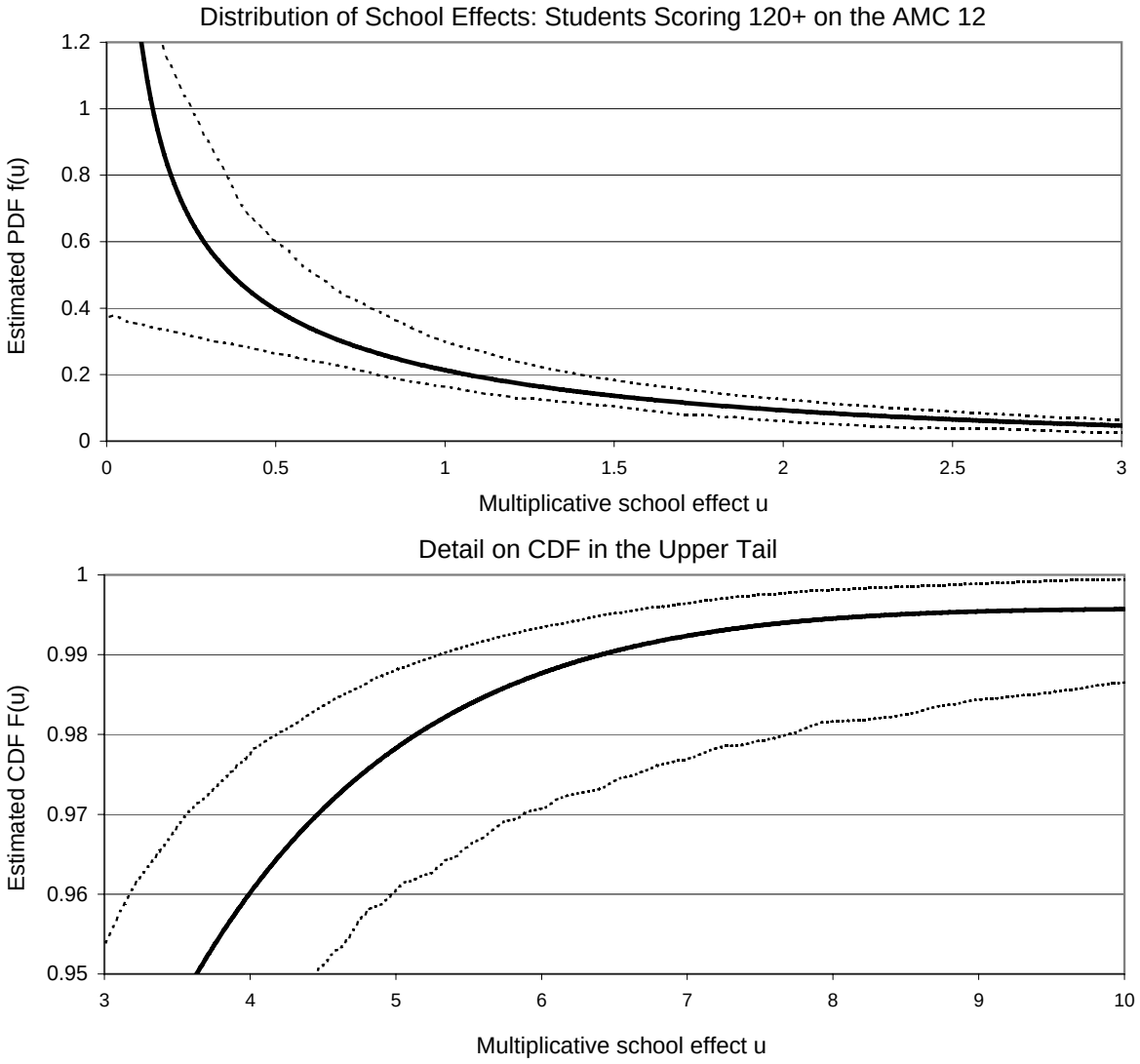

Figure 3: Distribution of school effects at a very high achievement level: students scoring 120+ on the AMC 12

average chance of scoring 120 on the AMC 12 (38% of ZIP codes are estimated to produce such students at less than one-quarter of the average rate.) There is also a very thick upper tail of ZIP codes where many more students reach this threshold.

We conclude that these features of our earlier estimates seem to be robust to moving to a full national sample. Of course, part of the reason for this is due to the nature of the data and estimation. Many ZIP codes that feed into schools that do not offer the AMC contests would be expected to have few high scorers given their demographics. This limits the information they provide to the estimation.

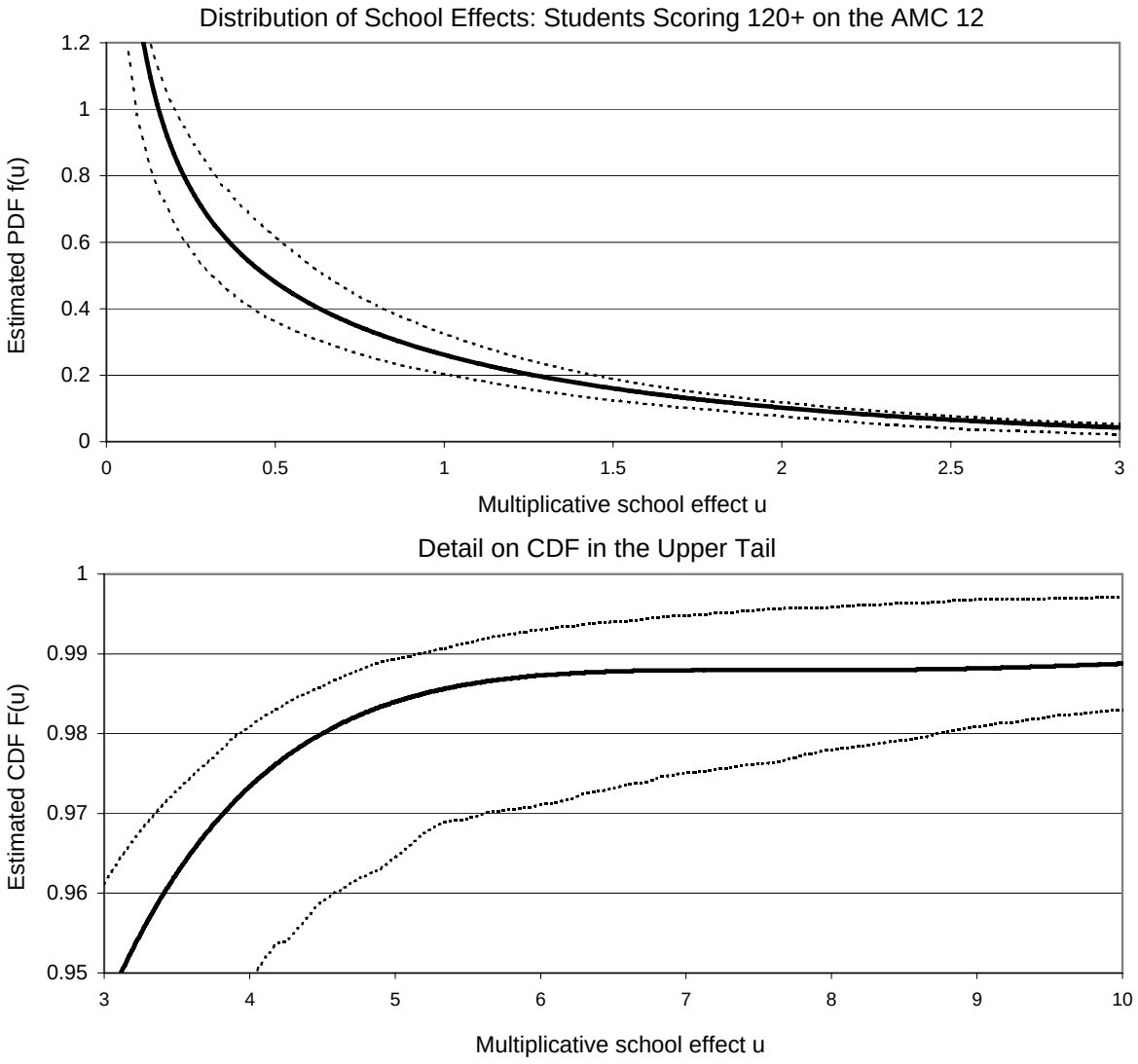

Figure 4: Distribution of ZIP code effects: students scoring 120+ on the AMC 12

6 High Scoring Students on the SAT

In this section we examine how likely different schools are to produce students with very high SAT scores. The primary motivation for this is that differences in the rate at which different schools produce students with high AMC scores will reflect both differences in educational programs and the self-selection of high-ability students into particular schools. Such selection effects should have similar effects on SAT performance. But one would expect that SAT performance will be less affected by differences in schools' educational programs: we presume that most schools that offer the AMC tests offer math and English courses that provide good coverage of SAT material. Hence, comparing estimates obtained from AMC and SAT data may provide insights on selection versus educational effects.

6.1 Data

Our source of data on students with high SAT scores are the announced lists of students who were named "candidates" for the U.S. Presidential Scholars Program (PSP). Being named as a PSP candidate can be roughly thought of as indicating that a student was among the twenty highest scoring male high school seniors or the twenty highest-scoring female high school seniors in their home state on the SAT Math + Critical Reading combined score.[30] Each top twenty is extended in the case of ties. In California, many more than 40 students score a perfect 1600 on the SAT, so being a PSP candidate is an indicator for having a perfect SAT score. Scores at or near 1600 are required in several other large states, but the cutoff is much lower in small and low-performing states. We obtained the full list of 2,752 PSP candidates for 2007 from the U.S. Department of Education website. PSP schools were linked to NCES data based on school name and student location.[31] To make results comparable we will carry out our school-level analyses on the subset of schools that participated in the 2007 AMC contests. This subsample includes 1,593 of the PSP candidates.[32]

6.2 Demographic patterns

The regression in the fifth column of Table 2 is another school-level regression run on our sample of nonmagnet, noncharter, coeducational public schools that offer the AMC ex-

[30] Students can also qualify via a high ACT score.

[31] We were able to match the schools attended by 2,520 of the 2,752 PSP candidates to the NCES data.

[32] The fact that at least 58% of PSP candidates attend schools that offer the AMC contests provides another datapoint suggesting that the majority of the high-achieving math students in the country probably attend schools offering the AMC.

ams using the number of 2007 PSP candidates at the school as the dependent variable.[33] The most important observation for our purposes is that the estimated α parameter from the negative binomial regression (0.48) is substantially smaller than the corresponding parameter (1.02) from the AMC 12 analysis. This indicates that there is less unobserved heterogeneity in the rates at which schools produce PSP candidates.

The demographic findings are in many ways similar to those concerning high AMC scorers: parental education matters a lot; median income remains negatively related to high achievement; and there are fewer high-achieving students in schools with more students qualifying for the free lunch program. The ethnic-group effects differ somewhat between the AMC- and PSP-based regressions. The strength of the Asian American effect is smaller in the PSP regression; and we now find fewer high-achieving students in schools with a larger Hispanic population. These differences may reflect that we have switched to a measure that also includes English test scores and requires less knowledge of math.

6.3 Distributions of unobserved heterogeneity across schools

Figure 5 presents an estimated distribution of school effects from PSP candidate counts comparable to our earlier estimates based on high AMC scorers.[34] Again, the top panel shows an estimated PDF for u's between 0 and 3 and the bottom panel provides a magnified view of the CDF for high values of u. The dashed lines are 95% confidence bands.

The most noteworthy finding is that the thick upper tail we found in the AMC data is not present in the PSP data. The estimated CDF reaches 0.995 at a point where the CDF estimated from the AMC data is just 0.97. Unless the selection into schools of students with high math ability differs from selection of students who score well on the SAT, this suggests that the thick upper tail in the AMC data is probably not due primarily to unobserved differences in student ability. Instead, our leading conjecture for what causes the thick upper tail is that there are very few schools, even among the very high quality schools that produce PSP candidates, that teach math at the level will result in many students doing well on the AMC 12.

The estimated density in the top panel also shows fewer schools with very low values of u – a feature that is very different from the point estimates from regressions examining

[33] Our PSP regressions use an additional control variable: *AwardRatio* is the ratio of the number of PSP candidates from the state to the number of public and private high school graduates in each state. This is intended to control for dramatic differences in how hard it is to be a PSP candidate in different states. For example, Wyoming, which has a population of 550,000, had 43 PSP candidates in 2010, whereas Michigan, with a population of about 10 million, had just 47.

[34] Coefficient estimates for the demographic variables are presented in the last column of the Table 2.

students scoring 120+ on the AMC 12. But the confidence bands are fairly wide at the lower end of the density.

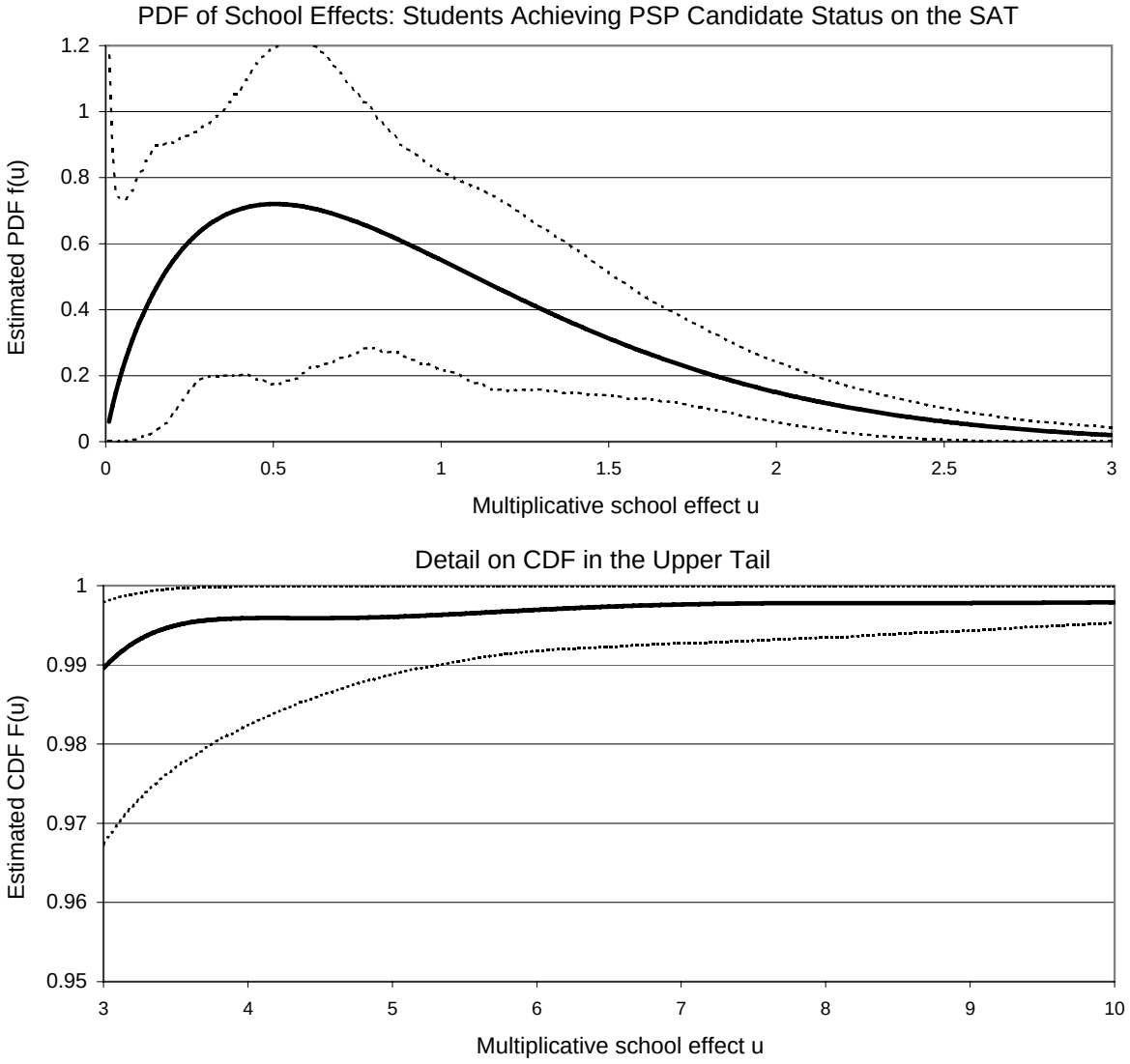

Figure 5: Distributions of school effects: Presidential Scholar Candidates

7 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 very far from being evenly distributed across the U.S.

There are very strong demographic predictors of high achievement as one might expect: areas where there are many highly educated parents and schools with many Asian-American students are much more likely to produce high-achieving students. Other patterns, however, are not necessarily what one would have expected. In the full sample of ZIP codes there are more high-scorers in high income areas. But once we restrict to the sample of schools offering the AMC, the median income in a community is negatively correlated with the likelihood that students will reach high levels of math achievement (and similarly negatively correlated with the number of students with very high SAT scores).

In multiple branches of the education literature there are results that have been roughly interpreted as implying that there is only limited variation across schools in value-added apart (perhaps) from variation that is systematically related to socioeconomic factors. Our results suggest that things look very different when one examines how likely schools are to produce very high-achieving students rather than on how schools' effects on average test scores. Our results suggest that there is a lot of variation among seemingly similar schools. The most notable feature of this variation is a thick uppper tail of schools in which students are many times more likely to reach high achievement levels than are students in the typical school with similar demographics. This thick upper tail is present in all of our analyses of students with high AMC scores, but is not present when we look at where students with high SAT scores are coming from. This contrast suggests that that the thick tail is not because of the self-selection of high-ability students into a particular subset of schools. We suggest that a potential explanation is that almost all schools see it as their responsibility to provide English and math courses that cover material necessary to do well on the SATs, whereas there is much less uniformity in whether schools encourage gifted students to develop more advanced problem solving skills and reach the higher level of mastery of high school mathematics needed to do well on the AMC.

Relative to the literature on the gender gap in mathematics, our comparison of school effects relevant to girls suggests that schools are perhaps even more important for girls: we estimate that the 99th percentile high school in our sample is producting high-scoring girls at more than ten times the rate of an average school with comparable demographics. We also note that there are many low-performing schools that will only very rarely have girls reach the AMC performance levels we have studied.

As we noted in the introduction, there are many ways in which one would like to do more than we can do with our data. It would be valuable to provide more evidence to separate out how much of the concentration of high-achieving students that we observe is due to

differences in school value added vs. differences in the selection of high-ability students into different public schools, to correlate differences in school policies with differences in outcomes, and to identify causal effects.

Our results suggest that the high-achieving math students we see today in U.S. high schools may be just a small fraction of the number of students who have the potential to reach such levels. This should probably not be surprising – the U.S. is far behind many other countries in the fraction of students who achieve very high scores on internationally administered tests. Our finding that there appears to be a lot of variation across schools with similar demographics could be seen as hopeful: the number of high-achieving students would increase substantially if low-achieving schools could be brought up to average; and upper-tail schools might have programs that could be emulated to produce even larger improvements.

Appendix

In this appendix we provide additional details on the estimation and simulation results assessing the performance of the estimator.

A.1 Estimation Methodology

For any fixed N our model of unobserved heterogeneity implies

$$\text{Prob}\{y_i = k|\tilde{z}_i\} = \int_0^\infty e^{-\left(e^{\tilde{z}_i\beta}u_i\right)} \frac{\left(e^{\tilde{z}_i\beta}u_i\right)^k}{k!} u_i^\alpha e^{-u_i} \left(\sum_{j=0}^{N} g_j L_j^{(\alpha)}(u_i)\right) du_i.$$

The estimated density will integrate to one if and only if $g_0 = 1/\Gamma(1+\alpha)$. We therefore set g_0 equal to this value and estimate β, α, and $g_1, \ldots, g_N$. The expression in Proposition 1 was used to compute the likelihood. The orthogonal polynomial term in the density estimation will take on negative and positive values for some parameter values (and in practice the integral defining the likelihood will sometimes be larger if the "density" is made negative in some places to allow it to be larger in others). In our estimation we have therefore chosen to add a penalty function of $\log\left(\int_0^\infty \max(\hat{f}(x), 0)dx\right)$ to the likelihood function for parameter values that do not generate valid densities.[35] The objective function is not globally concave, so we ran our estimation routines from a large number of starting values for the g_i.

We now give a derivation of the formula in Proposition 1. Let the density $f(x)$ be represented as $f(x) = x^\alpha e^{-x} \sum_{j=0}^\infty g_j L_j^{(\alpha)}(x)$.[36]. The distribution of y_i is then described

[35] 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 $\tilde{f}(x)$. One could impose nonnegativity as a numerical constraint in the estimation, but in practice we found that our optimization routine did not work well with this constraint and often ended with likelihoods much lower than those that would result from modifying "densities" that took on slightly negative values in the manner described above.

[36] The formula for the cdf can be derived as

$$\begin{aligned} F(x) &= \int_0^x \left(t^\alpha e^{-t} \sum_{j=0}^\infty \left(g_j \sum_{l=0}^{j} (-1)^l \binom{j+\alpha}{j-l} \frac{t^l}{l!}\right)\right) dt \\ &= \sum_{j=0}^\infty \left(g_j \sum_{l=0}^{j} (-1)^l \frac{\binom{j+\alpha}{j-l}}{l!} \left(\int_0^x t^{\alpha+l} e^{-t} dt\right)\right) \\ &= \sum_{j=0}^\infty \left(g_j \sum_{l=0}^{j} (-1)^l \frac{\binom{j+\alpha}{j-l}}{l!} \gamma(\alpha+l+1, x)\right), \end{aligned}$$

where $\gamma(.)$ is the lower incomplete gamma function.

by

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

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

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

To evaluate, first note that we have 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

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

We can then substitute these formulas into the formula for y_i:

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

Laguerre polynomials are orthogonal with

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

so that the formula for y_i simplifies to

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

This completes the proof.

Note that for an unrestricted parameter vector we would have

$$\int_0^\infty x^\alpha e^{-x} \sum_{j=0}^{\infty} g_j L_j^{(\alpha)}(x)\, dx = g_0 \Gamma(\alpha+1).$$

Accordingly we set g_0 so that this value is equal to one. To identify the parameters in the full semiparametric model we also need to impose the normalization that $E(u|z) = 1$. This can also be imposed by additionally restricting the parameter g_1 to be equal to $\alpha/\Gamma(\alpha+2)$. For finite N, however, it is not necessary to impose this for identification and in practice we do not impose the constraint and simply renormalize the estimated distribution by dividing by its expectation after the estimation stage.

A.2 Simulation Results

In this section we present some Monte Carlo estimates to illustrate how well the procedure described above works in some circumstances that may be roughly similar to the data in our application.

The simulations implemented our estimation procedure on datasets created by drawing each z_i from a uniform distribution with support $[0, i]$; 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 Poisson or negative binomial regression of the count of AMC 12 high-scorers on school- and school ZIP-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.

The estimated coefficients $\hat{\beta}$ on the observed characteristics are fairly precise and show almost no bias. Table 4 presents some summary statistics on the estimates for simulations

with $N = 8$ Laguerre polynomials.[37] 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.

	True	Mean and SD of estimated coefficients		
Variable	Coeffs.	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 $exponential(1)$ distribution, $lognormal(1, \frac{1}{3})$ distribution, and $uniform[0, 2]$ distribution with 2,500 simulated observations. Mean estimates across 1,000 simulated datasets shown; standard deviations in parentheses.

Table 4: Estimated coefficients on observed characteristics in simulations

Table 5 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 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 5 are median ISE across 1,000 simulation runs.

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

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

True distribution of u	Median ISE for various models				
	$N=0$	$N=2$	$N=4$	$N=6$	$N=8$
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 $exponential(1)$ distribution, $lognormal(1, \frac{1}{3})$ distribution, and $uniform[0,2]$ distribution with 2,500 simulated observations. Median ISE across 1,000 simulated datasets shown, varying the number of Laguerre polynomials.

Table 5: Goodness of fit of estimated distributions of unobserved heterogeneity in simulations: median MISE for various models and true distributions

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.

The lognormal distribution does not fit as well when $N = 0$, as one would expect: there is no α that gives an ISE of less than 0.0107. Larger N make it theoretically possible to fit the distribution much better (the best fit distributions have ISE 0.00756, 0.00210, 0.00014, and 0.00002 for $N = 2, 4, 6$, and 8), 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.[38] The best fit is obtained for $N = 6$, where the median ISE is 45% lower than the median ISE for the worst fit of $N = 2$.

Figure 6 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 method captures some important features of the distribution. The steep slope of the estimates at 0 and 2, and the double-peaked shape of the distributions in the range $[0, 2]$, allow the estimated functions to bound much of the estimated density in the correct

[38] The minimum possible ISE's are 0.0877, 0.0456, 0.0397, 0.0273, 0.0269 for $N = 0, 2, 4, 6, 8$.

support region.

A.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 $\{\tilde{y}_{ij}, \tilde{z}_{ij}\}_{i=1}^{2,051}$, 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 2,051 (with replacement) from the set of covariates $\mathbf{z}$ listed in Table 2. We also draw a random sample $\tilde{\mathbf{u}}_j$ of size 2,051 from the CDF $\hat{F}$, which we estimated using the procedure in Section 4 on the true dataset. For each $i = 1, ..., 2,051$, 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 $2,051$ (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.[39] 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 estimation and the center of each range of starting values for which the resulting likelihood

[39] 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.

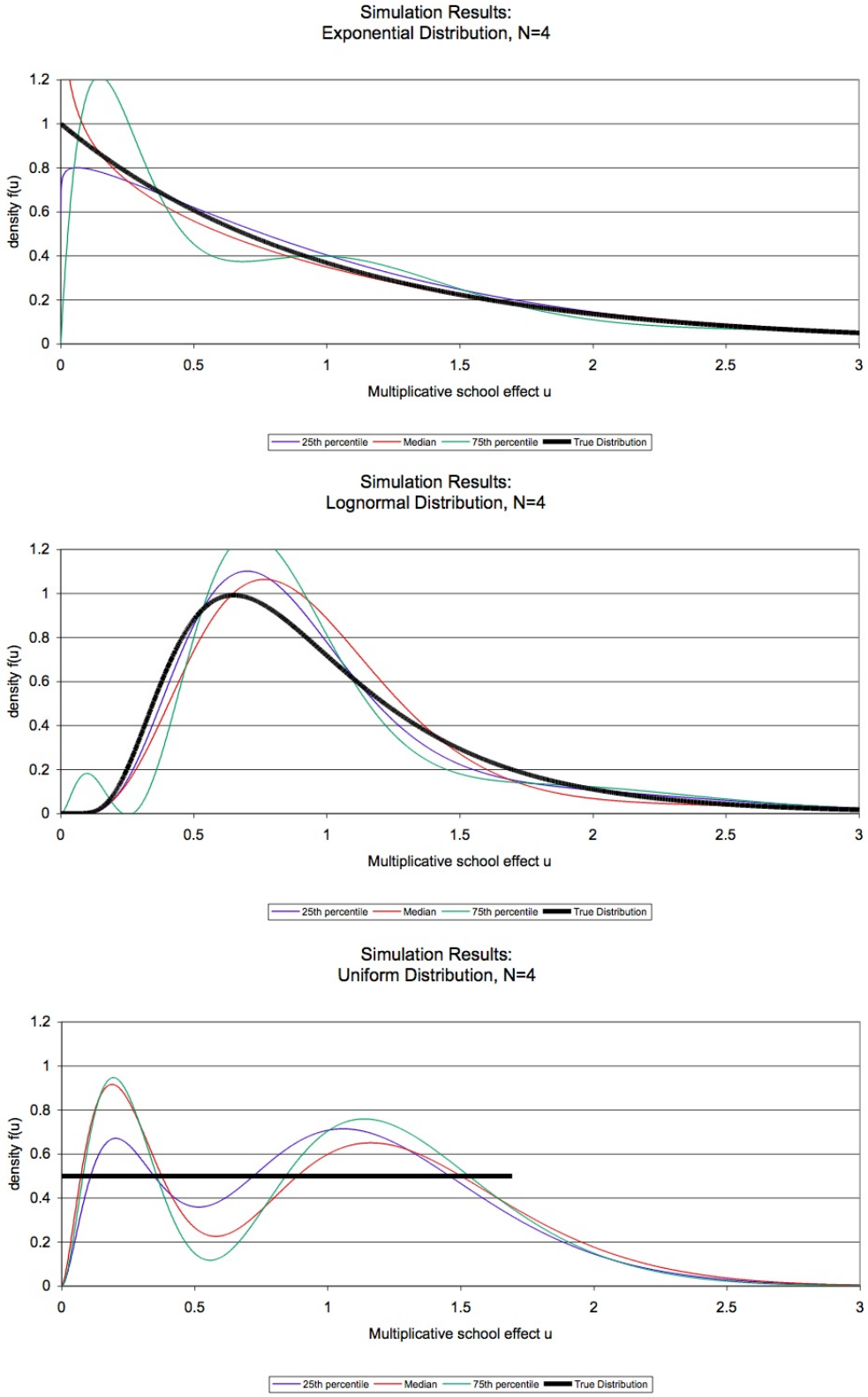

Figure 6: Actual vs. Estimated Distributions: 25^{th}, 50^{th}, and 75^{th} percentile fits in simulations

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.

References

Aaronson, Daniel, Lisa Barrow, and William Sander (2007). "Teachers and Student Achievement in the Chicago Public High Schools," *Journal of Labor Economics* 25, 95-135.

Abdulkadiroglu, Atila, Joshua Angrist, and Parag Pathak (2011): "The Elite Illusion: Achievement Effects at Boston and New York Exam Schools," mimeo, MIT.

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.

Brännäs, Kurt and Gunnar Rosenqvist (1994): "Semiparametric Estimation of Heterogeneous Count Data Models," *European Journal of Operational Research* 76, 247-258.

Brody, Linda, and Carol Mills (2005):, "Talent Search Research: What Have We Learned?," *High Ability Studies* 16(1), 97-111.

Brown, Charles, and Mary Corcoran (1997): "Sex-Based Differences in School Content and the Male-Female Wage Gap," *Journal of Labor Economics* 15(3), 431-465.

Bui, Sa, Steven G. Craig, and Scott Imberman (2011): "Is Gifted Education a Bright Idea: Assessing the Impact of Gifted and Talented Programs on Achievement," mimeo, University of Houston.

Cameron, A. Colin and Per Johansson (1997): "Count Data Regression Using Series Expansions: With Applications," *Journal of Applied Econometrics* 12, 203-223.

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.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Schanzenbach, and Danny Yagan (2011): "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," *Quarterly Journal of Economics* 126 (4), 1593-1660.

Coleman, James S. (1966): "Equality of Educational Opportunity Study (EEOS)." ICPSR06389-v3. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Colton, David, Heinz W. Engl, Alfred K. Louis, Joyce R. McLaughlin, and William Rundell (Eds.) (2000): *Surveys on Solution Methods for Inverse Problems*. Wien; New York: Springer.

Dee, Thomas, and Brian Jacob (2009): "The Impact of No Child Left Behind on Student Achievement," National Bureau of Economic Research Working Paper 15531.

Dobbie, Will and Roland G. Fryer, Jr. (2011): "Exam High Schools and Academic Achievement: Evidence from New York City," NBER Working Paper 17286.

Dobbie, Will and Roland G. Fryer, Jr. (2009): "Are High Quality Schools Enough to Close the Achievement Gap? Evidence from a Social Experiment in Harlem," NBER Working Paper 15473.

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

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

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.

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

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

Guo, Jie Q. and Pravin K. Trivedi (2002): "Flexible Parametric Models for Long-Tailed Patent Count Distributions," *Oxford Bulletin of Economics and Statistics* 64, 63-82.

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.

Heckman, J. and B. Singer (1984): "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica* 52(2), 271-320.

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

Hoxby, Caroline M. (2002): "School Choice and School Productivity (or Could School

Choice be a Tide That Lifts All Boats?)," NBER Working Paper 8873.

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

Kane, Thomas J., Jonah Rockoff, and Douglas O. Staiger (2006): "What Does Certification Tell Us About Teacher Effectiveness? Evidence From New York City," NBER Working Paper 12155.

Kane, Thomas J. and Douglas O. Staiger (2002): "The Promise and Pitfalls of Using Imprecise School Accountability Measures," *The Journal of Economic Perspectives* 16 (4), 91-114.

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.

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

McCaffrey, Daniel F., Daniel Koretz, J. R. Lockwood, and Laura S. Hamilton (2004): *Evaluating Value-Added Models for Teacher Accountability.* Santa Monica, CA: RAND Corporation, available at http://www.rand.org/pubs/monographs/MG158.

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.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain (2005): "Teachers, Schools, and Academic Achievement," *Econometrica* 73 (2), 417-458.

Rothstein, Jesse (2005): "SAT Scores, High Schools, and Collegiate Performance Predictions." Forthcoming, Rethinking Admissions for a New Millenium: Moving Past the SAT for Social Diversity and Academic Excellence, Joseph A. Soares, ed., Teacher's College Press.

Simar, L. (1976): "Maximum Likelihood Estimation of a Compound Poisson Process," *The Annals of Statistics* 4, 1200-1209.

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.

Wasserman, Larry (2006): "Density Estimation." *All of Nonparametric Statistics*. New York; London: Springer.

Winkelmann, Rainer (2008): *Econometric Analysis of Count Data*. Berlin: Springer.