Do Schools Matter for High Math Achievement? Evidence from the American Mathematics Competitions
Glenn Ellison and Ashley Swanson
Online Appendix

1 Appendix Tables

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

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

2 Proofs

Proof of Proposition 1

It is standard that under model 1 $Y_i \sim NB\left(\frac{1}{\alpha}, \frac{\alpha e^{\mathbf{X_i}\boldsymbol{\beta}}}{1+\alpha e^{\mathbf{X_i}\boldsymbol{\beta}}}\right)$ and under model 2 $Y_i \sim NB\left(\frac{\lambda(\mathbf{X_i})}{g(\mathbf{X_i})}, 1-e^{-g(\mathbf{X_i})}\right)$. (See Boswell and Patil (1970) or Karlin (1966, p. 345).) Hence, the distributions in the two models are identical if

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

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

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

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

Proof of Proposition 2

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

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

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

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

Proof of Proposition 3

Suppose the Y_{it} are generated as described. Then using Proposition 2 and

iterated expectations over t we have

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

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

$$\mathrm{Var}(Y_{it}|\mathbf{X_i}) = e^{\mathbf{X_i}\boldsymbol{\beta}} + e^{2\mathbf{X_i}\boldsymbol{\beta}}\left(\alpha_p\alpha_u + \alpha_p + \alpha_u\right).$$

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

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

The variance is is a little more complicated. We have

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

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

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

Plugging back into the formula for the variance we find

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

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

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

The first equation can hold for nonnegative (α_u, α_p) only if $\alpha_u \in [0, \alpha]$. Given any such α_u the first equation will hold for an unique α_p: $\alpha_p(\alpha_u) \equiv \frac{\alpha - \alpha_u}{1+\alpha_u}$. Given

this value for α_p we have $\alpha_p + \alpha_p\alpha_u = \alpha - \alpha_u$ so the second equation becomes $\overline{\alpha} = \alpha_u + \frac{1}{2}(\alpha - \alpha_u)$ which is true for $\alpha_u = 2\overline{\alpha} - \alpha$. The formula for α_p follows by substitution. $\square$

Proof of Proposition 4

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

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

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

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

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

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

and the series expansion

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

The former implies that

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

and the latter implies that

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

Substituting these formulas into the formula for y_i gives

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

Laguerre polynomials are orthogonal with

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

Using this, the the formula for y_i simplifies to

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

This completes the proof. □

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

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

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

3 Bootstrap Procedure

We obtained standard errors for our semiparametric estimates and confidence bands for the distribution of unobserved heterogeneity using both parametric and nonparametric bootstrapping procedures. In each iteration j of the bootstrap, we generate a simulated dataset $\{\tilde{y}_{ij}, \tilde{\mathbf{z}}_\mathbf{ij}\}_{i=1}^{1,984}$, then estimate the parame-

ters $\tilde{\alpha}_j, \tilde{g}_{j1}, ..., \tilde{g}_{jN}, \tilde{\boldsymbol{\beta}}_j$ using the semiparametric estimation procedure described in Section IV. 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

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

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 $\tilde{\alpha}_j, \tilde{g}_{j1}, ..., \tilde{g}_{jN}, \tilde{\boldsymbol{\beta}}_j$. 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 IV for the production of AMC high-scorers and in Section V for the production of SAT high-scorers.

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

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 $(\tilde{\mathbf{y}}^{NP}_j, \tilde{\mathbf{z}}_j)$ of size $1,984$ (with replacement) from the set of outcomes $\mathbf{y}$ and covariates $\mathbf{z}$, then estimate $\tilde{\alpha}^{NP}_j, \tilde{g}^{NP}_{j1}, ..., \tilde{g}^{NP}_{jN}, \tilde{\boldsymbol{\beta}}^{NP}_j$ on the simulated dataset $(\tilde{\mathbf{y}}^{NP}_j, \tilde{\mathbf{z}}_j)$. As in the semiparametric estimation on our full sample, the results of each bootstrap estimation may depend on the starting values chosen; in our results, we present those estimates for which the likelihood is highest after trying numerous starting values.[57] We begin each bootstrap by running a trial bootstrap of 20 simulations for several candidate starting values: those resulting in the highest likelihood in the full sample estimation and the center of each range of starting values for which the resulting likelihood is close to that of the best starting values. We then use the

[57] In practice, we used $\boldsymbol{\beta}$ 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.

values that provide the highest average log-likelihood in the trial bootstrap as the starting values in the full bootstrap.

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

4 Simulations

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

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

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

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

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

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

Table 6—: Estimated coefficients on observed characteristics in simulations

Table 7 provides some statistics on how well the model was able to estimate the distribution of unobserved heterogeneity. The rows correspond to the distribution from which the u's were drawn. The columns correspond to the number N of Laguerre polynomials used in the estimations. The metric used to measure performance is integrated squared error (ISE) – if the estimated density function from simulation run i is $\hat{f}_i(x)$, where the true data generation process has unobserved heterogeneity from distribution $f(x)$, the ISE of that estimated density is $\int_0^\infty (\hat{f}_i(x) - f(x))^2 dx$. The values in Table 7 are median ISE across 1,000 simulation runs.

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

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

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

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

The lognormal distribution does not fit as well when $N = 0$. This should be expected: the lognormal is not a member of the parametric family we are estimating and indeed no matter what α is estimated the ISE cannot possibly be below 0.0107. Larger N make it theoretically possible to fit the distribution

much better (the parameter vectors that give distributions closest to the true lognormal have ISEs of 0.00756, 0.00210, 0.00014, and 0.00002 for $N = 2, 4, 6,$ and 8 respectively), but again there is the offsetting effect that there is more scope for overfitting. The tradeoff between the two effects results in fairly similar fits across the range of N. The median ISE is smallest for the $N = 2$ model.

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

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

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

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

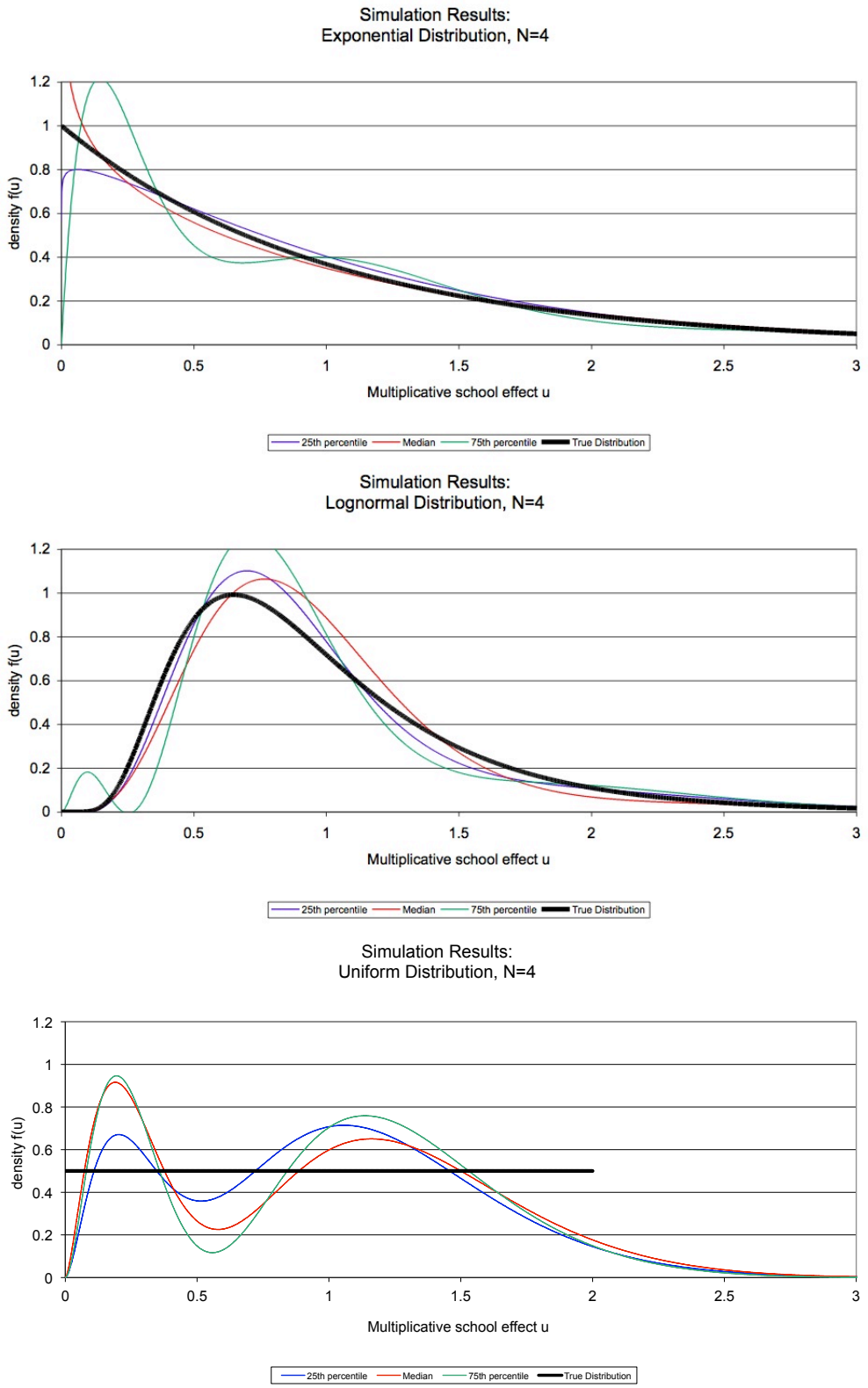

Figure 5. : Actual vs. Estimated Distributions: 25^{th}, 50^{th}, and 75^{th} percentile fits in simulations