1. Introduction

While improving public school education has been an empirical concern of parents, teachers, researchers, and policymakers for decades, a challenge has been the debate over the balance between increasing financial resources or pressing schools to improve efficiency. This has led to a multi-pronged policy approach in the United States (US), including both increased public-school spending - real per-pupil expenditures in public education increased from $7,000 in 1980 to $14,000 in 2015 (Baron, 2019) - and increased public school accountability - for example, the No Child Left Behind Act of 2001 (NCLB; Public Law 107-110). Nonetheless, student academic performance in the US continues to lag other Organization for Economic Co-operation and Development (OECD) countries despite spending more per pupil (Grosskopf et al., 2014). This suggests inefficiency in US public schools, where a lack of competitive market forces may allow it to persist. Consequently, econometrics production models that account for the existence of inefficiency are required, and this paper leverages the stochastic frontier literature (due to Aigner at al. 1977 and Meeusen and van den Broeck, 1977) to estimate and perform inference on inefficiency measures for public middle schools (serving grades 6-8) in New York City from 2014 to 2016. The nearest neighbors to our research are three recent stochastic frontier analyses of US public schools: Chakraborty et al. (2001), Kang and Greene (2001) and Grosskopf et al. (2014). Our research adds to this literature by applying a more flexible production specification (Greene 2005a, b) and modern inference techniques (Horrace, 2005; Flores-Lagunes et al., 2007), applied to data from the largest and one of the most diverse public-school systems in the country.

Public schools in New York City (NYC) enroll over 1.1 million students in more than 1,700 schools, of which over 200,000 are in middle school grades (grades 6 through 8) in more than 500 schools. The city's size and diversity provide a unique backdrop for a school efficiency study, because it has many schools (the primary unit of observation) that operate under a common set of regulations, funding mechanisms, and procedures, reducing the potential for heterogeneity bias due to differences in the economic and policy environment. Moreover, understanding school inefficiency in this environment is of great importance as 72.8% of students in NYC public schools are from economically disadvantaged backgrounds, a characteristic often negatively associated with educational attainment (Hanushek and Luque, 2003; Kirjavainen, 2012). To this end, we construct a balanced panel of 425 public middle schools that operate from 2012 to 2016 to estimate each school's technical inefficiency for the cohorts of students in grade 8 between the 2014 and 2016 academic years (AY). We begin with a school-level educational production function that measures output during middle school as the gains in mean students' test scores in Math and English Language Arts (ELA) between grade 5 (in the spring semester before students enter middle school grades) and grade 8 (in the last spring semester of middle school). We use gains in testing outcomes to address concerns that produced outputs (e.g., proficiency rates or mean test scores) are a result of student quality (selection into middle schools) rather than school efficiency. Our production function, then, also includes inputs that broadly fit into three groups - student characteristics, teacher characteristics, and school characteristics - in order to provide estimates of and to control for the marginal effects of other features of the middle school environment.

Aside from being the first stochastic frontier analysis of NYC public schools, to the best of our knowledge this paper is the first to apply the "true fixed effect stochastic frontier model" of Greene (2005a, b) to US school production.[1] This model is highly flexible, because it accounts for both persistent

[1] Kirjavainen (2012) is the only other education paper that applies Greene's model but to Finnish secondary schools.

(time-invariant) and transient (time-varying) inefficiency shocks. For example, Chakraborty et al. (2001) estimate only persistent inefficiency in a cross-section of Utah public schools. Kang and Greene (2001) estimate only transient inefficiency in an upstate NY public school district. Grosskopf et al. (2014) estimate only persistent inefficiency in public districts in Texas. We find that both persistent and transient inefficiency are present in NYC middle school production and ignoring either component is an empirical mistake.

In addition to improved flexibility of our specification relative to others, our paper considers different measures of transient inefficiency and uses inferential techniques that offer policymakers a methodology to determine groups of schools that are on the efficient frontier. In particular, parametric stochastic frontier models only yield a truncated (below zero) normal distribution of inefficiency conditional on the production function residual for each school. The most common approach to attain point estimates of school-level inefficiency is then to calculate the means of these conditional distributions (Jondrow et al., 1982) and rank them. However, the mean of a positive and continuous random variable can never be zero, so these point estimates can never identify efficient (inefficiency equal to zero) schools.[2] Therefore, in addition to calculating the means of these truncated normal distributions for each school, we calculate their modes as a point estimate of school-level efficiency (Jondrow et al., 1982). Since the truncated normal distribution for each school has a mode at zero inefficiency with positive probability, the mode measure allows for efficiency ties, producing a group of

[2] An exception in the stochastic frontier literature is the Laplace model of Horrace and Parmeter (2018), which yields conditional distributions with a probability mass at zero inefficiency.

firms that are on the efficient frontier. We also "salvage" the conditional mean point estimate using the inferential techniques in Horrace (2005) and Flores-Lagunes et al. (2007), which may be used to select a subset of schools that are efficient at the 95% level. We compare the cardinality of the set of mode-zero schools to the cardinality of the selected subset based on Horrace (2005).[3]

In the absence of frontier-based analyses, many studies estimate school (and teacher) effectiveness using value-added models (Ladd and Walsh, 2002; Meyer, 1997). We note these techniques are different in both purpose and form from the models we use here. Beginning with purpose, value-added models typically aim to identify the benefits of educational inputs (for example, if value-added increases when a policy is implemented) or the underlying quality of an education-producing unit (i.e., school or teacher), thus largely ignoring transient technical inefficiency. In fact, one of the major controversies of using value-added models for high-stakes public policy decisions stems from the assumption that deviations from each school's (or teacher's) fixed effect[4] may provide evidence that estimates are unstable (Koedel et al., 2015; Schochet and Chiang, 2013).[5] The true fixed effect stochastic frontier model allows for a portion of annual deviations to reflect transient inefficiencies in education production (perhaps, for example, related to effort or changes in curriculum) and to estimate the size of transient inefficiency for each unit. Then, in terms of difference in form, traditional value-added

[3] Mizala et al. (2002) proposed an approach for salvaging the conditional mean point-estimate. The divide production units into four quadrants using an efficiency-achievement matrix and treating those in the first quadrant as efficient. However, the approach is *ad hoc*, and is no substitute for a proper inference procedure.

[4] Some use random effects to estimate value-added, but this is relatively rare in the value-added literature.

[5] Another major controversy stems from bias that results from non-random student selection into schools (Angrist, et al., 2017; Ladd and Walsh, 2002).

models estimate the value-added of a unit as deviations from the conditional mean, while in our model we use the regression equation to develop an efficiency frontier. Using our probability statement technique, then, we can estimate the likelihood that individual units or groups of units operate on this efficiency frontier in a given observation year (or not). Conversely, value-added methods require decisionmakers to designate *ad hoc* cut-offs to assign policy levers, perhaps flagging high-value-added units for rewards or low-value-added units for penalty. Taken together, we believe the true fixed effect stochastic frontier model can address some of the major controversies that surround the use of value-added models or previous stochastic frontier techniques used for education policymaking, in part because the model is intended to identify inefficiency rather than quality, and in part because it separates persistent from transient inefficiencies, which allows for better targeting of policy levers towards each form of inefficiency.

In short, we find that student composition of a school is more predictive of production in ELA, while the teacher composition of a school is more predictive of Math production, which is consistent with conventional wisdom that ELA achievement is more reflective of home and individual characteristics, and Math achievement is more reflective of classroom characteristics (Bryk and Raudenbush, 1988). Second, by separating persistent technical inefficiency from transient technical inefficiency, we are able to show that both sources of inefficiency harm the productivity of middle schools in NYC (the conditional means of both sources range from about one-half to a whole standard deviation, depending on subject considered and estimator used). Third, we offer evidence that both efficient and inefficient schools operate in all five boroughs of NYC, suggesting school inefficiency is geographically dispersed and dispersed across schools serving high and low performing students. Fourth, by separating inefficiency from the error term (under our set of distributional assumptions), decisionmakers are better able to assess the extent to which declining exam performance during middle school is due to inefficiency as opposed to statistical noise. Finally, we offer policymakers a pair of actionable decision rules that are methodologically rigorous and reflect true performance of schools, both derived from the true fixed effects model, including application of the conditional mode estimator to identify when schools operate efficiently or the more rigorous Horrace (2005) probabilities to identify a subset of the best.

The rest of the paper is organized as follows. The next section presents the econometric model and reviews the stochastic frontier literature as it relates to research in educational inefficiency. Section 3 discusses the data. Section 4 presents the empirical results. Section 5 concludes.

2. Stochastic Frontier Models in Education Efficiency

Stochastic frontier analysis (SFA) is an econometric technique to estimate a production function while accounting for statistical noise and inefficiency. A highly flexible specification for panel data is due to Greene (2005a, b), who considers the linear production function:

$$y_{it} = \alpha + x_{it}'\beta + v_{it} - u_{it} - w_i, \quad i = 1, \dots, n, \quad t = 1, \dots T, \tag{1}$$

where $u_{it} \geq 0$ is a random effect representing transient (time-varying) inefficiency of the i^{th} school in period t, $w_i \geq 0$ is a fixed- (or random-) effect, and v_{it} is the usual mean-zero random error term (or regression noise). The variable y_{it} is productive output (e.g., student proficiency rates, average test scores, or gains in test scores). The x_{it} is a vector of productive inputs (e.g., financial and nonfinancial resources, student characteristics and baseline performance, teacher quality and experience, principal quality, and other productive inputs), β is an unknown vector of marginal products, and α is an unknown constant. Assuming w_i is fixed, let unobserved heterogeneity be $\alpha_i = \alpha - w_i$, leading to the Greene

(2005a, b) true fixed-effect stochastic frontier model.[6] In general, w_i captures all forms of time-invariant unobserved heterogeneity. Nonetheless, the SFA literature refers to w_i as "persistent technical inefficiency," and we will follow the same practice in what follows. Our empirical focus is characterizing and making inferences on u_{it}.

Identification of the model requires mutual independence of the random error components and the inputs over *i* and *t*. Since the mean of u_{it} (conditional on inputs) is non-zero, identification also requires parametric distributional assumptions on the random error components, typically $v_{it} \sim N(0, \sigma_v^2)$ with $u_{it} \sim |N(0, \sigma_u^2)|$ (half normal) or u_{it} distributed exponential with variance σ_u^2.[7] Then a within- or first-difference transformation of the model and maximum likelihood estimation leads to consistent estimates of α, β, σ_u^2, σ_v^2 (as T or $n \to \infty$), and the MLE residuals can be used to consistently estimate α_i (as $T \to \infty$). A consistent estimate of α is the maximum of the estimated α_i, and a consistent estimate of persistent inefficiency (w_i) is the difference between the estimated α and each estimated α_i. The parametric assumptions (whether *u* is half normal or exponential) imply that the distribution of transient inefficiency (u) conditional on $\varepsilon_{it} = v_{it} - u_{it}$ is a truncated (at zero) normal distribution parameterized in terms of the estimates of $\sigma_u^2, \sigma_v^2,$ and *T* with the regression residuals (e_{it}, say), substituted for errors ε_{it} (Aigner et al., 1977).

---

[6] Assuming fixed *w* allows identification of the model even when *w* is correlated with *x*, the usual panel data result.

[7] Other distributions for *u* have been proposed, such as truncated normal (Stevenson, 1980), gamma (Greene, 1980a,b), uniform and half Cauchy distribution (Nguyen, 2010) and truncated Laplace (Horrace and Parmeter, 2018). Kumbhakar and Lovell (2015) show that the choice of distribution most likely does not affect the relative ranking of estimated firm-level inefficiency.

Point estimation of firm-level (transient) inefficiency proceeds by calculating moments of the truncated normal distribution of *u* conditional on $\varepsilon_{it} = e_{it}$. Jondrow et al. (1982) provide formulae for the conditional expectation, $E(u|\varepsilon_{it} = e_{it})$, and the conditional mode, $M(u|\varepsilon_{it} = e_{it})$, which are reproduced in the Appendix. The conditional mean is more commonly employed in empirical exercises as a point estimate for inefficiency but has the shortcomings that it is always positive and that the probability of ties across i is zero.[8] That is, no firm is on the efficient frontier and there are never ties in the efficiency scores. On the other hand, the conditional mode, allows for ties at zero.[9] We calculate both point estimates of transient inefficiency in our application, but suggest that the oft-ignored conditional mode may be a more useful point estimate for policymakers. That is, the mode determines a group of schools to be on the efficient frontier, so policy prescriptions can be made for the group of schools that are under-preforming or to reward schools operating efficiently. This phenomenon is illustrated in Figure 1, which plots the conditional mean and mode for the Normal-Half Normal (NHN) specification and for the Normal-Exponential (NE) specification for continuous values of ε_{it} with $\sigma_u^2 = \sigma_v^2 = 1$ and $\alpha = \beta = 0$.

Selecting the schools with conditional mode equal to zero is a useful policy tool, but it is not a decision rule grounded in statistical theory, so we also appeal to the selection rule in Flores-Lagunes et al. (2007) based on the efficiency probabilities of Horrace (2005), which we briefly describe here and for

---

[8] This is an empirical fact to anyone familiar with the empirical literature. It is likely due to economist's preferences for conditional expectations.

[9] To see this, consider a $N(\mu, \sigma^2)$ density truncated at zero. If $\mu > 0$, the mode is positive, otherwise it is zero.

which we provide more details in the Appendix. Given the *n* truncated normal conditional (transient) inefficiency distributions of *u* and given a specific time period *t*, we follow Horrace (2005) to characterize transient inefficiency as the probability that school *i*'s draw of *u* is the smallest in any period *t*,

$$\pi_{it} = \Pr\,(u_{it} < u_{jt},\ j \neq i | \varepsilon_{1t}, \dots, \varepsilon_{nt}).$$

These are within-sample, relative "efficiency probabilities." Then one may estimate the probabilities by substituting $\varepsilon_{it} = e_{it}$ above and use the estimated efficiency probabilities to select a subset of schools that contains the unknown efficient school at a prespecified confidence level (e.g., 95%), following Flores-Lagunes et al (2007).[10] Let the population rankings of the unknown efficiency probabilities be,

$$\pi_{[n]t} > \pi_{[n-1]t} > \cdots > \pi_{[1]t},$$

and let the sample rankings of the estimated probabilities, $\hat{\pi}_{it}$, be

$$\hat{\pi}_{(n)t} > \hat{\pi}_{(n-1)t} > \cdots > \hat{\pi}_{(1)t},$$

where $[i] \neq (i)$ in general. Then, the Flores-Lagunes et al. (2007) procedure is to sum the estimated probabilities, $\hat{\pi}_{it}$, from largest to smallest until the sum is at least 0.95. Then the school indices in the sum represent a "subset of the best schools," containing the unknown best school, $i = [n]$, with probability at least 95%. Equivalently, the school indices in the subset of the best cannot be distinguished and are all on the within-sample efficient frontier (in a statistical sense). If the subset of the best is a singleton, then there is only one efficient school, $[n] = (n)$. The subset could contain all *n* schools, so all schools are on the frontier. The lower the cardinality of the subset, the sharper the statistical inference on [*n*].

---

[10] We do not show how to do this, so the reader is referred to Horrace (2005) and Flores-Lagunes et al. (2007).

Education researchers have adopted SFA to estimate production frontiers and to analyze school inefficiency, including: universities (Dolton et al. 2003, Gronberg et al. 2012, Stevens 2005, and Zoghbi et al. 2013); school districts (Chakraborty et al. 2001, Grosskopf et al. 2014, and Kang and Greene 2001); and primary and middle schools (Garcia-Diaz et al. 2016, Kirjavainen 2012, Pereira and Moreira 2011; Salas-Velasco 2019).[11] Only a few of these studies focus on inefficiency in US public school education. Chakraborty et al. (2001) set T = 1 and w = 0 in (1) to measure the inefficiency of public education in Utah. Kang and Greene (2001) set w = 0 in (1) to analyze technical inefficiency in an upstate NY public school district from 1989 to 1993. Grosskopf et al. (2014) set T = 1 and w = 0 in (1) to analyze data from 965 public school districts in Texas. In all these papers, the only estimate of US school-level inefficiency considered is the conditional mean, $E(u|\varepsilon_{it} = e_{it})$, and none of these papers consider inference over the identification of efficient and inefficient schools in any meaningful way.

Compared to the other, earlier models, the "true fixed effect" model relaxes the assumption that technical inefficiency must be time invariant and allows for unobserved school heterogeneity. Unlike Greene (2005a, b), however, we estimate the model using marginal maximum simulated likelihood estimation (MMSLE), proposed by Belotti and Ilardi (2018).[12] The maximum likelihood dummy variable estimation originally proposed by Greene (2005a, b) suffers from an incidental parameter problem, resulting in inconsistent estimates of σ_u^2 and σ_v^2.[13] MMSLE addresses the incidental parameter problem

[11] Surveys of SFA in education are Worthington (2001), Johnes (2004) and De Witte and López-Torres (2017).

[12] This estimation is available on Stata in command *sftfe*.

[13] More detailed explanation of the incidental parameter problem can be found in Neyman and Scott (1948) and Lancaster (2000).

by treating the marginal likelihood function as an expectation with respect to the change of residuals and estimates variances through simulation. MMSLE also allows for consideration of both normal-half normal and normal-exponential distribution assumptions for the technical inefficiency parameter, u_i.[14]

3. Data

We use data from the New York State Education Department (NYSED) and New York City Departments of Education (NYC DOE) to construct a balanced panel of education outputs (test score gains) and education inputs (student, teacher, and school characteristics) for cohorts of NYC public school students that completed middle school between AY 2014 and AY 2016. Specifically, we use school-level data from the NYS School Report Cards (SRC), which contains information on school enrollments by grade, student demographics, and teacher characteristics in every NYS public school. We merge SRC data to aggregated student data that summarizes the mean gains in Math and English Language Arts (ELA) test scores between grades 5 and 8 for each cohort in every school as well as mean characteristics of those test takers.[15] The resulting panel contains 425 public middle schools in NYC, excluding charter schools and schools that open, close, or otherwise are missing data during our sample period. The schools are scattered across all five NYC boroughs, including 133 in Brooklyn, 115 in the Bronx, 84 in Manhattan, 80 in Queens, and 13 in Staten Island.

[14] Chen et al. (2014) proposes an alternative using marginal maximum likelihood estimation (MMLE), which utilizes closed skew normal distributions properties (González-Farías et al., 2004) to derive closed-form expressions of the marginal likelihood function to address the incidental parameter problem.

[15] In the following, unless specified, we use test-takers and students interchangeably.

3.1 Educational Outcomes

We construct cohort-level measures of normalized test score gains to measure schools' education production. We use test scores on annual standardized exams implemented by the New York State Testing Program (NYSTP), which administrates state-wide mathematics (Math) and English language arts (ELA) tests to students from grade 3 to grade 8 in compliance with the standards of the NCLB Act and, later, the "Every Student Succeeds Act (ESSA) of 2015" (Public Law 114-95, 2015).[16]

Following common practice in education economics research, we normalize student test performance across grades and years as standardized z-scores with a mean of zero and a standard deviation of one for each grade and year, thus pegging performance to the citywide mean for each cohort. The standardized exams are administered in the second half of each academic year (usually in April or May), so we calculate z-score changes ("gains") between grade 5 and grade 8 to reflect education production during the middle school period (which spans grade 6 to grade 8).[17] Thus, for example, if a student is at precisely the citywide mean for students in grade 5 in AY 2012 and one standard deviation above average in grade 8 in AY 2015, their gain score takes a value of one (1). This has implications for interpretations of the marginal products in equation 1. For example, if β equals 0.5 for a variable in x_{it}, such as the share of students with limited English proficiency, then increasing this share of students from

[16] More information can be found on https://www.schools.nyc.gov/learning/in-our-classrooms/testing.

[17] We also use specifications that treat grade 8 z-scores as the output, either with baseline performance in grade 5 included as a student characteristic or without that additional variable. The first of these models are akin to value-added models and produce similar results to those presented in this paper. The second do not control for baseline performance (an all-too-common practice in previous SFA research), so some estimates differ because they reflect both marginal effects and uncontrolled student quality.

0 to 1 increases average gains in test scores by one-half of a standard deviation. For our main sample, we restrict each cohort to those students who take both the Math and the ELA standardized exams in both grade 5 and grade 8 to limit the extent to which the composition of a cohort changes by students transferring into and out of NYC schools and the bias that results from nonrandom selection into the testing population by exam (such as students taking one exam but not the other due to expected performance). By including only students with complete exam data in each cohort, we ensure that the mean cohort-level gain scores reflect true changes in performance over time for the same students, rather than changes in the composition of test takers.[18]

3.2 Educational Inputs

Following Grosskopf et al. (2014), we include school, teacher, and test-taker characteristics among our educational inputs. Column one of Table 1 lists input variables included in this study. Test-taker characteristics include sociodemographic information, such as share of the cohort by race/ethnicity (white, black, Hispanic, Asian, or multiracial), gender, with limited English proficiency, with disabilities, and from economically disadvantaged households. Teacher characteristics include the number of teachers per one hundred students, and teacher quality measures, such as the share of teachers with a master's degree or greater, teaching without valid certification, out of certification, and who have more than three years of experiences. School characteristics include the share of classes taught by teachers

[18] To test the sensitivity of our results to cohort restrictions, we relax the sample constraints to keep students with either complete (grade 5 and 8) Math or ELA exams (rather than both subjects). Results are substantively similar (in magnitude and direction) to the main results reported and are available from the authors upon request.

without certification, the average number of classes per one hundred students, the number of staff (excluding teachers) per one hundred pupils, and the number of principal and assistant principals per one hundred students.

The second column of Table 1 reports citywide summary statistics of the educational inputs. Hispanic students are the largest racial/ethnic group in NYC, accounting for nearly half of students in the average middle school, followed by black students at 34.7%. More than three-fourths of students in the average NYC public middle school are economically disadvantaged, and roughly 17% are students with disabilities. We also report summary statistics by borough in columns 2-6 of Table 1. The share of white students accounts for only 3.88% in middle schools in the Bronx, but nearly half for the schools in Staten Island. Compared with other boroughs, schools in the Bronx also have the largest share of students from economically disadvantaged backgrounds (83.94%) and with limited English proficiency (8.83%). In terms of teacher and school inputs, middle schools in the Bronx have the highest share of teachers out of certificate (20.62%) and without valid certification (1.58%). Schools in Staten Island is at the other end of the spectrum, having the lowest mean shares of students from economically disadvantaged background (58.52%) or with limited English proficiency (1.52%). The share of teachers with master or higher degrees (66.58%) and with three or more years of experience (94.12%) are also the highest in Staten Island. We note, as well, that performance varies across districts, with the mean grade 8 Math and ELA z-scores 16% and 20% of a standard deviation below average for schools in the Bronx, but 25% and 21% of a standard deviation above average for schools in Staten Island. Average middle school gains in test performance also vary by district, but not to the same degree; the borough with the smallest gains is the Bronx with 7% and 1% of a standard deviation gains in Math and ELA, respectively, and the borough with the greatest gains is Manhattan with 16% and 11% of a standard deviation gains in those two subjects.[19]

4. Results

Estimates from the "true" fixed-effect stochastic frontier model in equation 1 are shown in in Table 2. We only present estimates for Math (column 2) and ELA (columns 3) scores using normal-exponential and normal-half normal specifications of the model, respectively. The normal-half model for Math and the normal-exponential model for ELA did not converge, so estimates are not presented.

4.1 Marginal Effect of Education Inputs

Columns 2 and 3 of Table 2 contain the marginal effects for improvements in Math and ELA scores, respectively. Generally speaking, we find that improvements in Math scores are largely uncorrelated with test-taker and school characteristics, while teacher characteristics are important. Improvement in ELA scores are largely due to student characteristics.

Beginning with the marginal effects of test-taker characteristics, we find none of the student characteristics are correlated with middle school Math gains at the 95% significance level (though "share multiracial" is positively and "share limited English proficient" is negatively correlated with Math gains with p-values less than 0.1). Conversely, share female, Asian, and limited English proficiency are all positively correlated with ELA gains (while other test taker characteristics are not). For example, an

[19] All gain scores calculated as the difference between grades 8 and 5 mean performance. At first blush, it is counterintuitive that gain scores are above 0 for all boroughs, but we note that this reflects that students entering the district in middle school are lower performing than those enrolled and who take the exams in both grades.

increase in the share of a cohort who is female from none to all (0 to 1) is associated with greater gains during middle school of nearly one-fifth of a standard deviation (0.190). Put differently, a 10 percentage-point increase in the female share of students is correlated with 1.90 percent of a standard deviation greater increases in gain scores. Similarly, 10 percentage-point increases in share of a cohort who are Asian or with limited English proficiency increase ELA gains by 4.53% and 3.85% of a standard deviation, respectively.

Unlike test taker characteristics, we find teacher characteristics are more strongly correlated with Math performance gains than ELA. As the number of teachers per 100 pupils increases by 1, Math gains increase by 3.49% of a standard deviation. As the share of teachers with at least three years of experience increases by 10 percentage-points, Math gains increase by 6.62% of a standard deviation. Perhaps surprisingly, share of teachers with master's or doctorate degrees is negatively associated with gains in Math (a 10 percentage-point increase is linked with 6.62% of a standard deviation decrease in Math gains) and share of teachers without certification are positively associated (a 10 percentage-point increase is linked with 13.64% of a standard deviation greater gains). None of these teacher characteristics are correlated with ELA gains.

School characteristics appear to matter little for education production in both subjects, because none of the school characteristics are significantly correlated with gains in middle school Math or ELA performance at the 95% level (though the number of professional staff per 100 pupils is positively correlated with Math gains at the 90% level and the number of classes per 100 pupils is positively correlated with ELA gains at the 90% level).

4.2 Persistent Technical Inefficiency Estimates

After controlling for production inputs, Figure 3 summarizes the distribution of our estimates of Persistent Technical Inefficiency (PTI) by borough and by test subject (Math or ELA). That is, the figure plots the empirical distribution of our estimates of $w_i = \alpha - \alpha_i$. The rectangular boxes show the medians, 25th, and 75th percentiles of Persistent Technical Inefficiency (PTI) for each subject and borough. The lower and upper whiskers below and above each box are the percentiles that are 1.5 times the interquartile range below and above the 25th and 75th percentiles, respectively, for each subject and borough. The dots are individual estimates of PTI for schools outside the whisker percentiles: the most and least persistently efficient schools in the sample. For example, there are two dots at PTI = 0, indicating that the persistently efficient ELA school is in Brooklyn and the persistently efficient Math school in the Bronx. It also appears that there is a second Bronx school that is very close to the efficient frontier in the Math test. In general, we find that the interquartile ranges of PTI are largely higher (and, perhaps, wider) in the Bronx, Brooklyn, and Manhattan than in Queens and Staten Island. Differences in estimated PTI are less stark for ELA, but it does seem they are slightly higher in the Bronx than elsewhere. Of greater note, perhaps, is that the distributions of inefficiency across the NYC's boroughs are not so large as to reflect a "tale of two cities" - there are schools in the Bronx that are estimated to have low PTI as well as schools in Staten Island with moderate to moderately high estimated PTI. We note that direct comparisons across the two subjects should be avoided, because the educational production functions for Math and ELA are estimated separately with different distributional assumptions on the transient inefficiency component, *u*.

We report the mean and standard error of Persistent Technical Inefficiency (PTI) in Table 3. Consistent with Figure 3, the Bronx has the highest mean PTI: 1.08 for Math and 0.58 for ELA, both of which are significantly higher than the average citywide PTI. In other words, over the period the Bronx is persistently about one standard deviation below the efficient frontier of normalized test score improvements in Math and about a half a standard deviation below the frontier in ELA. Conversely, Staten Island has the smallest PTI for Math and ELA (0.82 and 0.45 for Math and ELA, respectively), and differences from the citywide mean are statistically significant. Under our modelling assumptions, this implies that schools in the Bronx persistently operate less efficiently on average than those in Staten Island (or Queens, for the matter). Given that these schools also serve the lowest performing students, as shown in Table 1, the results suggest that PTI increased the student achievement gap across boroughs during this period.

Do schools with large Persistent Technical Inefficiency (PTI) in Math also have large PTI in ELA? Figure 4 presents a scatterplot of PTI in Math against PTI in ELA in all years with a linear fit line superimposed (the slope of the line is 1.13, with a t-statistic of 16.35). A Spearman test, comparing school ranks in Math PTI and ELA PTI, finds a positive (0.6169) and significant statistic (p-value = 0.0000), suggesting a strong monotonic relationship between PTI in Math and PTI in ELA.

4.3 Transient Technical Inefficiency

Table 4 shows summary statistics of each school's Transient Technical Inefficiency (TTI) with plotted distributions for Math and ELA shown in Figures 4 and 5, respectively. Remember, all that these models admit is the truncated (at zero) normal distribution of TTI conditional on the residual values of 425 school in each of 3 years. Here we point estimate (summarize) these conditional distributions for each of the 425 * 3 = 1,275 school-years using their conditional means and modes (and later the conditional probability that each school is efficient) as described in section 2 and the Appendix. The first row of Table 4 contains summary statistics for the conditional mean of the Math TTI distributions for all schools in all years. For example, the mean of the conditional mean point estimates of TTI for Math is 0.115. That is, conditional on the residuals, we expect that Math TTI is 0.115 (3rd column) for all schools and years. Thus, on average TTI reduces improvements in Math scores by 0.115 standard deviations in the sample, which is comparable in magnitude to the mean gains in Math scores during this period (0.12 standard deviations citywide as reported in Table 1). Put differently, the grade 8 student achievement gap in Math for schools in the Bronx and Staten Island is approximately 0.41 standard deviations (as indicated in Table 1); the mean citywide TTI is 28% the size of that gap.

The first row of Table 4 contains other statistics for the conditional mean estimates of Math TTI as well. For example, the observation with the minimal conditional mean point estimate for Math TTI has a value of 0.022, implying that it is 0.022 standard deviations below the efficient frontier. That is, based on the conditional mean estimates, the most efficient school-year in the sample for Math TTI is *inefficient in expectation*. Therefore, the conditional mean point estimate of TTI is made relative to an out-of-sample standard (a theoretical best school whose TTI distribution can be described as a Dirac delta at $u = 0$). The first row of Table 4 also reports the 25th, 50th and 75th percentiles of the conditional means of Math TTI distributions, as well as the maximal point estimate, which implies that we expect the least efficient school-year in the sample to be 0.715 standard deviations below the (theoretical) efficiency frontier.

The second row in Table 4 summarizes the conditional mode point estimates of the Math TTI distributions. Compared to the conditional mean point estimates (first row), which are expectations, the conditional modes provide estimates of the most common (or likely) value of TTI for each observation. While the conditional mean is a measure of central tendency that can never equal zero for a non-negative u, the conditional mode can occur anywhere in the non-negative support of the truncated normal distributions that characterizes TTI. In particular we see in the second row of Table 4 that the average of the conditional mode point estimates is 0.045, which is considerably lower than the average of the conditional mean estimates (0.115) in the first row. We also see that the minimal estimate of the conditional mode is exactly zero (5th column). That is, for this school-year the most likely draw from its conditional distribution of TTI is $u = 0$, an efficient draw. Looking across the second row in Table 4, this is also true of the school at the 25th percentile (6th column) and the median school (7th column), meaning that at least half the schools in the sample are *likely* to be efficient (their conditional mode is on the frontier) even though they are appear inefficient in expectation (their conditional mean is not). While the conditional mean and the conditional mode of TTI summarize the truncated normal distributions in different ways, the mode has the added benefit of providing an *ad hoc* decision rule for selecting efficient schools: those with conditional modes equal to zero.[20] For example, in the last row of the table we see that the minimal value for the conditional mode of the ELA TTI is zero (5th column), as expected. However, the 25th percentile is positive 0.018 (6th column), implying that at least 75% of the observed school-years are unlikely to be efficient.

Finally, we note that in Table 4 the maximum conditional mean and the mode estimates appear to be the same for Math TTI, 0.715 (last column, first and second rows) and for ELA TTI, 0.339 (last column, third and fourth rows), but this equivalence is rounding error. Due to the nature of a normal distribution truncated at zero, the distribution's mean is always larger than its mode. For the maximal

[20] A statistically rigorous decision rule is based on the Horrace (2005) efficiency probabilities, and is considered in the sequel.

school-year observation in the last column of Table 4, however, the amount of truncation is so small that at three significant digits (0.715 for Math and 0.339 for ELA) its effect is negligible.[21]

Figures 5 and 6 plot the empirical distributions of our TTI estimates in Table 4 (note the axes scales for the two subjects differ). The conditional mean (red) and mode (blue) distributions in Figure 5 correspond to the summary statistics in rows 1 and 2 (respectively) of Table 4, while the distributions in Figure 6 correspond to the statistics in rows 3 and 4 of the table. The usefulness of the conditional mode as a decision rule for selecting efficient schools each year is clear. In Figure 5, the blue spike at zero indicates that more than 60% of the school-year observations in the sample are likely to be efficient in terms of their conditional distributions of Math TTI. In Figure 6 the blue spike indicates that about 19% of the school-year observations in the sample are likely to be efficient in terms of their conditional distributions of ELA TTI. Again, this is an *ad hoc* decision rule, but one that is easily understood by policymakers. What could a policymaker make of the red distributions of the conditional means in Figures 5 and 6? Not much compared to the blue distributions of the conditional modes in these figures.

In Table 5 we compare TTI by borough and academic year, reporting the percentage of schools with zero estimated TTI based on the conditional mode point estimate and our *ad hoc* decision rule. The table is self-explanatory. Staten Island has the highest percentage of schools operating efficiently (5th column) in Math (62%), followed by Queens (60%), the Bronx (59%), Manhattan (58%) and Brooklyn

[21] As with any truncated normal distribution with a very large mean equal to its mode (due to symmetry), the distribution is no longer symmetric after truncation at zero. That is, its new, post-truncation mean is necessarily larger than its mode, which is unchanged when the pre-truncated mean is positive. Moreover, the mean shifts further to the right as the amount of truncation increases.

(55%) in order. Conversely, Manhattan and Brooklyn have the highest percentage of transiently efficient schools (last column) in ELA (18%), followed by the Bronx (17%), Queens (13%), and then Staten Island (10%). Looking at the trend over time, we find middle schools in the Bronx show consistent improvements in the percentage of schools operating efficiently, while other boroughs do not have consistent improvements in efficiency over the period. The share of middle schools in the Bronx with zero TTI increases from 49% to 60% to 67% in Math and 11% to 16% to 24% in ELA. Schools in Queens, on the other hand, are less likely to operate with zero TTI in Math over time (71% to 65% to 45%) with no consistent trends in ELA (20% to 6% to 11%). All other boroughs also do not display consistent positive or negative trends in TTI.[22]

Figure 7 shows a weak but positive relationship between Math TTI and ELA TTI during our sample period (the slope of the line is 0.56 with a t-statistic of 12.62). The Spearman test statistics each year range between 0.2834 and 0.3312 and are statistically significant.

4.4 Efficiency Probabilities

As suggested previously, the above-described *ad hoc* rule to identify efficient school-year observations lacks statistical rigor. Therefore, we calculate school-level efficiency probabilities (Horrace, 2005) to identify the subset of schools that operate efficiently each year in terms of TTI (Flores-Lagunes et al., 2007). A more thorough discussion of the technique is contained in the Appendix, but as stated earlier, it uses the conditional truncated normal distribution of TTI for each school to calculate the probability

[22] While it is tempting to compare the magnitudes of TTI in Table 4 to the PTI in Table 3, the reader is reminded that PTI may also contain other sources of time-invariant unobserved heterogeneity, so comping TTI to PTI is ill-advised in general.

that each school is efficient in each year (has the smallest *u* conditional on the data), then selects a minimum cardinality subset of schools that contain the efficient school with at least 95% probability. This is a rigorous statistical decision rule, and we have two goals in the analyses that follow. The first is to calculate this minimal cardinality "subset of the best" schools in each year at the 95% level, and the second is to compare the cardinality of the subset of the best in each year to the cardinality of the subset of zero-mode schools in each year.

Table 6 contains the results. The first row of the table is for Math TTI in AY 2014. In 2014 there were 246 schools (3rd column) with conditional modes equal to zero. Since we have a balanced panel of 425 schools, this means that 58% percent of schools are efficient in 2014 based on our *ad hoc* decision rule, and this number corresponds to the 58% in the last row, 2nd column of Table 5. Call these 246 schools the "zero-mode subset" of efficient schools. The 4th column of Table 6 is the sum of the efficiency probabilities for the schools in the zero-mode subset. That is, the probability that the most efficient firm in the sample is contained in the zero-mode subset is 75.22% in 2014 for Math TTI. Put another way, the schools in the zero-mode subset are statistically indistinguishable from the unknown efficient school with probability 75.22%. Thus, the efficiency probabilities allow us to assign a confidence level to our *ad hoc* decision rule, however it is well below typical confidence levels like 90% or 95%. Nonetheless there is policy value in knowing the zero-mode subset. The second and third rows of Table 6 show that the cardinalities of the zero-mode subsets (3rd column) in AY 2015 and 2016 (259 and 231 schools respectively) are similar to AY 2014, as are the zero-mode probabilities (4th column), which are 76.10% and 73.09%, respectively. In sum, we are about 75% confident that about 60% of NYC schools are likely operating efficiently in terms of Math TTI.

Moving to the ELA test results in Table 6 (rows 4-6), we see that the cardinalities of the zero-mode subsets (3rd column) are 61, 75, 72 with probabilities (4th column) 32.30%, 37.97% and 40.77% in AY 2014, 2015 and 2016, respectively. These are much lower cardinalities and probabilities than the Math TTI results, and this is reflected in the smaller blue spike at zero in Figure 6 compared to Figure 5 (which have different scales). Why might this be the case? Is there something inherent in ELA education that lends itself to lower efficiencies relative to Math education? It is, in fact, a commonly found empirical phenomenon that ELA achievement is more reflective of home environment and individual characteristics, while Math performance is more responsive to classroom characteristics - a finding that is also consistent with theory about the pedagogy of ELA and Math instruction (Bryk and Raudenbush, 1988). It is noteworthy too, that our results in Table 2 also show test-taker characteristics are more predictive of greater ELA gains, while teacher characteristics are correlated with greater Math gains. Keep in mind, however, that we are assuming that the distribution of Math TTI is half-normal and that of ELA TTI is exponential, and this was driven by nonconvergence of the likelihood maximization in alternative specifications. Aside from this technical detail, it may simply be that mathematical standards for "correctness" are objective and those for the language arts are more subjective, so identifying "best practices" in ELA may be more difficult than in Math. There is a branch of SFA that attempts to explore the determinants of technical inefficiency (e.g., Cho and Schmidt, 2020). Perhaps such an analysis may be helpful here, but this is left for future research.

The 5th column of Table 6 contains the cardinality of the 95% minimal cardinality subset of the best (Flores-Lagunes et al., 2007), and in the first row we see that for Math TTI in 2014, 372 of our 425 middle schools were indistinguishable from the best middle school in the sample with 95% probability. Here, and in the other rows of the table, the zero-mode subset is always a proper subset of the subset of the best. (Whether this is a coincidence or not remains to be seen and is left for future research.) The implication is that even if the zero-mode decision rule is *ad hoc* and does not achieve usual confidence levels, at least it identifies a subset of schools that are contained in the subset of the best, as based on a rigorous statistical decision rule.

Looking down the 5th column across academic years for Math TTI, 369 to 374 (depending on year) of the 425 schools are statistically indistinguishable from efficient at the 95% level. For ELA, 326 to 339 of 425 schools are statistically indistinguishable from efficient. These are useful statistics that policymakers may use to determine which and how many schools to target when designing interventions that are intended to improve performance.

5. Conclusions

This study provides summaries of persistent and transient technical efficiency estimates for each of 425 NYC middle schools using recent advancements in stochastic frontier modeling. Using the "true" fixed effect stochastic frontier model to estimate gains in mathematics (Math) and English language arts (ELA), we find substantial variation in persistent technical inefficiency across the city and between boroughs. We note that while some boroughs have higher shares of persistently inefficient schools than others, there is a wide and overlapping distribution across each of the five boroughs in the city, suggesting school efficiency in NYC is not a "tale of two cities". Thus, while the mean Math and ELA persistent technical inefficiency in the city are 0.99 and 0.53 standard deviations, respectively - both larger than the student achievement gap between schools in the borough that enrolls the highest performing students (Staten Island) and schools in the borough that teaches the lowest performing students (the Bronx) - school inefficiency itself is widely distributed across the NYC's boroughs and schools. Still, to give a sense of magnitude of the results, if the city could find a way to remove persistent technical inefficiency in schools in the Bronx, for example, it would eliminate the achievement gap across boroughs (and, in fact, even overshoot the target).

We next produce estimates of transient technical inefficiency, using both a conditional mean and a conditional mode estimator. Under the conditional mode estimator and an *ad hoc* decision rule, we find around 58% of schools are transiently technically efficient in Math and 16% in ELA. We then apply a probability statement approach to offer rigorous inferential insights on which school-years are statistically on the efficient frontier, and which are very likely not. Based on the results of our "zero-mode subset" and the minimal cardinality subset of the best, the model can be used for both subjects to provide substantial information to decisionmakers on which schools likely did and did not operate efficiently each year. These are important distinctions for policymakers to be able to make; for example, the difference in the mean achievement gains for students attending a school-year observation in the zero-mode subset in ELA using the conditional mode is estimated to make 6.2 percent of a standard deviation greater gains than if that school were operating at the median level of inefficiency for ELA in that year (equivalent to about 15% of the gap between mean grade 8 achievement in Staten Island and the Bronx).

Another innovation of this study is the use of student-level academic performance data to estimate gains over time, which are then aggregated to the cohort-school-level to more accurately measure the education produced during the middle school years. These sorts of "gains models" are common in other education research but have not yet been used in stochastic frontier modelling. This innovation allows for improved estimates of the marginal effects of student, teacher, and school inputs on education production as well as a more compelling methodology for determining which schools are persistently efficient in each year.

Our results suggest that policymakers should more rigorously consider the role inefficiency plays in reducing education production in public schools. First, we identify which features (and types of features) in the school environment are beneficial or harmful to education production in middle schools, interestingly finding that student composition of a school is more important for the production of ELA gains, while teacher composition of a school is more important for the production of Math gains. These results are consistent with the conventional wisdom that ELA achievement is more responsive to home and individual characteristics and Math achievement is more responsive to classroom characteristics (Bryk and Raudenbush, 1988).

Second, by separating persistent technical inefficiency from transient technical inefficiency, we offer a methodology for school district administrators to separate the systemic features of a school that harm efficiency (such as, perhaps, building or principal quality) from those that change perennially (such as, perhaps, teacher effort or curriculum design). Third, by separating inefficiency from the error term (under the above-described distributional assumptions), decisionmakers are better able to assess the extent to which declining exam performance is due to inefficiency as opposed to statistical noise.

Fourth, for any of these considerations, arming policymakers with actionable decision rules that are methodologically rigorous and reflect true performance of schools is a tall ask of any statistical model. We believe that using the proposed true fixed effects model with either the conditional mode estimator to identify when schools operate efficiently or using the more rigorous Horrace (2005) probabilities present districts with useful tools to make high-leverage decisions such as bonus pay, promotions, intervention targeting, among others. The methods presented here can address some of the shortcomings of previous work estimating the effectiveness of schools, by offering estimates of inefficiency (rather than other quality measures) and separating persistent from transient inefficiencies. These advances may allow for better targeting of policy levers towards disincentivizing each form of inefficiency, with threat of dismissal or reorganization working towards reducing the former and with docked pay or performance incentives perhaps reducing the latter. This work motivates future efforts to estimate the effects of such policies on both forms of technical inefficiency.

References

Aigner, D., Lovell, C., & Schmidt, P. (1977). Formulation and Estimation of Stochastic Frontier Production Function Models. *Journal of Econometrics*, 6(1), 21-37.

Joshua D. Angrist, J.D., Hull, P.D., Pathak, P.A., & Walters, C.R. (2017). Leveraging Lotteries for School Value-Added: Testing and Estimation. *The Quarterly Journal of Economics*, 132(2), 871-919.

Baron, E. J. (2019). School Spending and Student Outcomes: Evidence from Revenue Limit Elections in Wisconsin. Available at SSRN 3430766.

Belotti, F., & Ilardi, G. (2018). Consistent Inference in Fixed-Effects Stochastic Frontier Models. *Journal of Econometrics*, 202(2), 161-177.

Bryk, A.S., & Raudenbush, S.W. (1988). Toward a More Appropriate Conceptualization of Research on School Effects: A Three-Level Hierarchical Linear Model. *American Journal of Education*, 97(1), 65-108.

Chakraborty, K., Biswas, B., & Lewis, W. C. (2001). Measurement of Technical Efficiency in Public Education: A Stochastic and Nonstochastic Production Function Approach. *Southern Economic Journal*, 889-905.

Chen, Y.-Y., Schmidt, P., & Wang, H.-J. (2014). Consistent Estimation of the Fixed Effects Stochastic Frontier Model. *Journal of Econometrics*, 181(2), 65-76.

Cho, C-K and P. Schmidt (2020). The Wrong Skew Problem in Stochastic Frontier Models when Inefficiency Depends on Environmental Variables. *Empirical Economics* 58, 2031-47.

De Witte, K., & López-Torres, L. (2017). Efficiency in Education: A Review of Literature and a Way Forward. *Journal of the Operational Research Society*, 68(4), 339-363.

Dolton, P., Marcenaro, O. D., & Navarro, L. (2003). The Effective Use of Student Time: A Stochastic Frontier Production Function Case Study. *Economics of Education Review*, 22(6), 547-560.

Flores-Lagunes A, Horrace WC, Schnier KE. (2007). Identifying Technically Efficient Fishing Vessels: A Non-Empty, Minimal Subset Approach. *Journal of Applied Econometrics*, 22, 729-745.

Garcia-Diaz, R., del Castillo, E., & Cabral, R. (2016). School Competition and Efficiency in Elementary Schools in Mexico. *International Journal of Educational Development*, 46, 23-34.

González-Farías, G., Domínguez-Molina, A., & Gupta, A. K. (2004). Additive Properties of Skew Normal Random Vectors. *Journal of Statistical Planning and Inference*, 126(2), 521-534.

Greene, W. H. (1980a). Maximum Likelihood Estimation of Econometric Frontier Functions. *Journal of Econometrics*, 13(1), 27-56.

Greene, W. H. (1980b). On the Estimation of a Flexible Frontier Production Model. *Journal of Econometrics*, 13(1), 101-115.

Greene, W. (2005a). Fixed and Random Effects in Stochastic Frontier Models. *Journal of Productivity Analysis*, 23(1), 7-32.

Greene, W. (2005b). Reconsidering Heterogeneity in Panel Data Estimators of the Stochastic Frontier Model. *Journal of Econometrics*, 126(2), 269-303.

Gronberg, T. J., Jansen, D. W., & Taylor, L. L. (2012). The Relative Efficiency of Charter Schools: A Cost Frontier Approach. *Economics of Education Review*, 31(2), 302-317.

Grosskopf, S., Hayes, K. J., & Taylor, L. L. (2014). Efficiency in Education: Research and Implications. *Applied Economic Perspectives and Policy*, 36(2), 175-210.

Hanushek, E. A., & Luque, J. A. (2003). Efficiency and Equity in Schools Around the World. *Economics of Education Review*, 22(5), 481-502.

Horrace, W. C. (2005). On Ranking and Selection from Independent Truncated Normal Distributions. *Journal of Econometrics*, 126(2), 335-354.

Horrace, W. C. & Richards-Shubik, S. (2012). A Monte Carlo Study of Ranked Efficiency Estimates from Frontier Models. *Journal of Productivity Analysis*, 38(2), 155-165.

Horrace WC, Richards-Shubik S, Wright IA. (2015) Expected Efficiency Ranks from Parametric Stochastic Frontier Models. *Empirical Economics*, 48, 829-848.

Horrace, W. C. & Parmeter, C. F. (2018). A Laplace Stochastic Frontier Model. *Econometric Reviews*, 37(3), 260-280.

Johnes, J. (2004). 16 Efficiency Measurement. *International Handbook on the Economics of Education*, 613.

Jondrow, J., Lovell, C. K., Materov, I. S., & Schmidt, P. (1982). On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model. *Journal of Econometrics*, 19(2), 233-238.

Kang, B. G., & Greene, K. V. (2002). The Effects of Monitoring and Competition on Public Education Outputs: A Stochastic Frontier Approach. *Public Finance Review*, 30(1), 3-26.

Kirjavainen, T. (2012). Efficiency of Finnish General Upper Secondary Schools: An Application of Stochastic Frontier Analysis with Panel Data. *Education Economics*, 20(4), 343-364.

Koedel, C., Mihaly, K., Rockoff, J.E. (2015). Value-Added Modeling: A Review. *Economics of Education Review*, 47, 180-195.

Kumbhakar, S. C., & Lovell, C. K. (2003). Stochastic Frontier Analysis. Cambridge university press.

Ladd, H.F., & Walsh, R.P. (2002). Implementing Value-Added Measures of School Effectiveness: Getting the Incentives Right. *Economics of Education Review*, 21, 1-17.

Lancaster, T. (2000). The Incidental Parameter Problem Since 1948. *Journal of Econometrics*, 95(2), 391-413.

Meeusen, W., & van Den Broeck, J. (1977). Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error. *International Economic Review*, 435-444.

Meyer, R.H. (1997). Value-Added Indicators of School Performance: A Primer. *Economics of Education Review*. 16(3), 283-301.

Mizala, A., Romaguera, P., & Farren, D. (2002). The Technical Efficiency of Schools in Chile. *Applied Economics*, 34(12), 1533-1552.

Neyman, J., & Scott, E. L. (1948). Consistent Estimates Based on Partially Consistent Observations. *Econometrica: Journal of the Econometric Society*, 1-32.

Nguyen, N. B. (2010). Estimation of Technical Efficiency in Stochastic Frontier Analysis (Doctoral dissertation, Bowling Green State University).

Pereira, M. C., & Moreira, S. (2007). A Stochastic Frontier Analysis of Secondary Education Output in Portugal. Available at SSRN 1398692.

Salas‐Velasco, M. (2020). Assessing the Performance of Spanish Secondary Education Institutions: Distinguishing Between Transient and Persistent Inefficiency, Separated from Heterogeneity. *The Manchester School*, 88(4), 531-555.

Schochet, P.Z., & Chiang, H.S. (2013). What Are Error Rates for Classifying Teacher and School Performance Using Value-Added Models? *Journal of Educational and Behavioral Statistics*, 38(2), 142-171.

Stevens, P. A. (2005). A Stochastic Frontier Analysis of English and Welsh Universities. *Education Economics,* 13(4), 355-374.

Stevenson, R. E. (1980). Likelihood Functions for Generalized Stochastic Frontier Estimation. *Journal of Econometrics*, 13(1), 57-66.

Worthington, A. C. (2001). An Empirical Survey of Frontier Efficiency Measurement Techniques in Education. *Education Economics*, 9(3), 245-268.

Zoghbi, A. C., Rocha, F., & Mattos, E. (2013). Education Production Efficiency: Evidence from Brazilian Universities. *Economic Modelling*, 31, 94-103.

Appendix

The Conditional Mean and Mode

When *v* is normal and *u* is half-normal, the model is Normal-Half Normal (NHN). When *u* is exponential, the model is Normal-Exponential (NE). Per Jondrow et al. (1982), the closed-form expressions of the conditional mean under normal-half normal and normal-exponential assumptions are:

$$E(u_{it}|\varepsilon_{it}, NHN) = \sigma_* \left[\frac{\phi\left(\frac{\varepsilon_{it}\lambda}{\sigma}\right)}{1 - \Phi\left(\frac{\varepsilon_{it}\lambda}{\sigma}\right)} - \left(\frac{\varepsilon_{it}\lambda}{\sigma}\right) \right], \qquad E(u_{it}|\varepsilon_{it}, NE) = \sigma_v \left[\frac{\phi(A)}{1 - \Phi(A)} - A \right]$$

where $\sigma^2 = \sigma_u^2 + \sigma_v^2, \sigma_*^2 = \sigma_u^2\sigma_v^2/(\sigma_u^2 + \sigma_v^2), \lambda = \frac{\sigma_u}{\sigma_v}$ and $A = \frac{\varepsilon_{it}}{\sigma_v} + \frac{\sigma_v}{\sigma_u}$. ϕ and Φ are the probability density function and cumulative distribution function of standard normal distribution. Estimates are formed by substituting the MMLE estimates for their population parameters into these formulae while setting $\varepsilon_{it} = e_{it}$. A less commonly employed estimator proposed by Jondrow et al. (1982) is the mode of the conditional distribution of $u_{it}|\ \varepsilon_{it}$, denoted as $M(u_{it}|\ \varepsilon_{it})$, to measure transient technical inefficiency. Under normal-half normal and normal-exponential distribution assumptions, the conditional mode estimator can be written as:

$$M(u_{it}|\varepsilon_{it}, NHN) = \begin{cases} -\varepsilon_{it}\left(\frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2}\right), & if\ \varepsilon_{it} \leq 0, \\ 0, & if\ \varepsilon_{it} > 0\,. \end{cases}$$

$$M(u_{it}|\varepsilon_{it}, NE) = \begin{cases} -\varepsilon_{it} - \frac{\sigma_v^2}{\sigma_u}, & if\ \varepsilon_{it} \leq -\frac{\sigma_v^2}{\sigma_u}, \\ 0, & if\ \varepsilon_{it} > -\frac{\sigma_v^2}{\sigma_u}\,. \end{cases}$$

The parametric forms of both conditional mean and conditional mode estimators under NHN and NE are functions of ε_{it}. To better understand the differences between the conditional mean and the conditional mode estimators, we standardize the standard errors σ_v and σ_u to one and plot their relationships with ε_{it} under NHN in Figure 1 and under NE in Figure 2. The figures show that both conditional mean and conditional mode estimators are monotonically decreasing with the regression residual. The conditional mode estimator, however, is always below the conditional mean estimate given the same residual. Moreover, when the residual surpasses a threshold (0 under NHN or $-\frac{\sigma_v^2}{\sigma_u}$ under NE), the conditional mode estimator takes a value of zero whereas the conditional mean estimator is positive and monotonically decreasing. This is intuitive - the more negative the regression residual, the farther the school is below that frontier and the more likely it is to be operating with large inefficiency. When the regression residual is large and positive, the school's estimated productivity is above the production frontier, suggesting the inefficiency is likely to be small. The difference between the estimators, then, is that, when above the threshold, the estimated TTI using the conditional mean estimator is small but still positive, whereas using the conditional mode estimator is zero. We use this conditional mode property to identify "zero-mode" schools that are likely to be operating efficiently.

Similar to the conditional mean estimator, the conditional mode estimator can be used to rank schools. However, unlike the conditional mean, the ranking allows for ties if more than one school is estimated to have zero TTI. Among schools with positive conditional mode estimates (non-zero estimated inefficiency), however, the order of the rankings is the same as from the conditional mean.

Conditional Efficiency Probabilities and the Subset of the Best Schools

While conditional mean estimates can be used to rank schools and conditional mode estimates can be used to find zero-mode efficient schools, neither estimate can produce joint probability statements on the relative ranking of the schools. To assess the reliability of the results and to draw inference on the efficiency rankings, we turn to the probability statement approach (Horrace, 2005; Flores-Lagunes et al., 2007; Horrace and Richards-Shubik, 2012; Horrace et al., 2015). Assuming independence of *u* over *i* and *t*, the probability of the event "school i is efficient at time t" is:

$$\pi_{it} = P\{\, u_{it} \leq u_{jt} \;\forall\, i \neq j \mid \varepsilon_{1t}, \dots, \varepsilon_{nt} \,\} = \int_0^{\infty} f_{u_{it}|\,\varepsilon_{it}}(u) \prod_{j\neq i}^{n} \left[1 - F_{u_{jt}|\,\varepsilon_{jt}}(u)\right] du,$$

where $f_{u_{it}|\,\varepsilon_{it}}(u)$ and $F_{u_{it}|\,\varepsilon_{it}}(u)$ are the probability density function and cumulative distribution function of $u_{it}|\,\varepsilon_{it}$, respectively. If *u* is half-normal with variance σ_u^2, then $u_{it}|\,\varepsilon_{it}$ is $N^+(-\frac{\varepsilon_{it}\sigma_u^2}{\sigma_v^2+\sigma_u^2}, \frac{\sigma_u^2\sigma_v^2}{\sigma_u^2+\sigma_v^2})$. If *u* is exponential, then $u_{it}|\,\varepsilon_{it}$ is $N^+(-\varepsilon_{it} + \sigma_v^2/\sigma_u, \sigma_v^2)$. To estimate the probabilities π_{it}, the regression residuals, e_{it}, are substituted into the above formulas for errors, ε_{it}. Then, given any subset of the *n* schools (like our zero-mode subset), we can assign a confidence level to the set containing the efficient school by summing the probabilities π_{it} for the schools in the set. Alternatively, let the population rankings of the unknown efficiency probabilities be,

$$\pi_{[n]t} > \pi_{[n-1]t} > \cdots > \pi_{[1]t},$$

and let the sample rankings of the estimated probabilities, $\hat{\pi}_{it}$, be

$$\hat{\pi}_{(n)t} > \hat{\pi}_{(n-1)t} > \cdots > \hat{\pi}_{(1)t},$$

where $[i] \neq (i)$ in general. We can determine a 95% minimal cardinality subset of the best school by summing the probabilities from the largest (*n*) to the smallest (1) until the sum is at least 0.95. Then, the school indices in the sum are “in contention for the best school” with probability at least 95% at time t. In other words, these schools cannot be statistically distinguished from the (unknown) best school in the population, $[n]$. For example, if $\hat{\pi}_{(n)t} > 0.95$, then the minimal cardinality subset is a singleton containing only the index (n), and the inference is very sharp. If $\hat{\pi}_{(n)t} < 0.95$, but $\hat{\pi}_{(n)t} + \hat{\pi}_{(n-1)t} > 0.95$ (say), then the minimal cardinality subset is $\{(n), (n-1)\}$. It is possible that the subset contains all n schools, $\{(n), (n-1), \dots, (1)\}$. This occurs when $\sum_{i=1}^{n-1} \hat{\pi}_{(i)t} < 0.95$ or equivalently when $\hat{\pi}_{(1)t} > 1 - 0.95$.