The 'Pupil' Factory:
Specialization and the Production of Human Capital in Schools*

Roland G. Fryer, Jr.
Harvard University and NBER

February 2016

Abstract

Starting in the 2013-2014 school year, I conducted a randomized field experiment in fifty traditional public elementary schools in Houston, Texas designed to test the potential productivity benefits of teacher specialization in schools. Treatment schools altered their schedules to have teachers specialize in a subset of subjects in which they have demonstrated relative strength (based on value-add measures and principal observations). The average impact of teacher specialization on student achievement is -0.042 standard deviations in math and -0.034 standard deviations in reading, per year. Students enrolled in special education and those with younger teachers demonstrated marked negative results. I argue that the results are consistent with a model in which the benefits of specialization driven by sorting teachers into a subset of subjects based on comparative advantage is outweighed by inefficient pedagogy due to having fewer interactions with each student. Consistent with this, specialized teachers report providing less attention to individual students (relative to non-specialized teachers), though other mechanisms are possible.

* I thank Terry Grier and Andrew Houlihan, for help and guidance in conducting the experiment. I am grateful to Will Dobbie, Richard Holden, Simon Jaeger, Andrei Shleifer, Jörg Spenkuch, and Chad Syverson for helpful comments and suggestions. Meghan Howard and Rucha Vankudre provided exceptional project management oversight. Tanaya Devi, Lisa Phillips, and Brecia Young provided excellent research assistance. Financial Support from Eli Broad and the Edlab's Advisory Group is gratefully acknowledged. Correspondence can be addressed to the author by email at rfryer@fas.harvard.edu. The usual caveat applies.

I. Introduction

Smith (1776) begins his analysis of the wealth of nations with the causes and consequences of the division of labor among workers.[1] Through his famous analysis of pin factories in 18th century England, Smith (1776) demonstrated the power of specialization in economics by arguing that in traditional production processes, factories would produce one pin per day per worker. Yet, by streamlining the eighteen-step process of pin production into nine individual tasks, the factory could produce 4,800 pins per worker.

Another striking example of the potential productivity gains from the division of labor is the assembly line approach to automobile production. In assembly line production, workers, machines and parts are sequentially organized and workers add parts to the machine as it moves from work station to work station. Henry Ford broke the assembly task into 84 discrete steps and trained workers to do just one step to increase his factories' productivity. This reduced the production time of a car from 12.5 hours to 93 minutes. Production figures compiled from the Model T Comprehensive Encyclopedia show that production before the assembly line was introduced in 1913 averaged 68,773 cars per year. In 1913, production increased to 170,211 cars in a year (McCalley, 1989).

In almost every modern industry, comparative advantage is used to maximize productivity. Goods produced by individual craftsmen have become so rare due to their relatively high cost that they now represent a niche "artisanal" market.

The basic economics is intuitive. Specializing in the production of a subset of the tasks necessary to produce a final output allows workers to gain efficiency in that task. Smith identifies three main channels through which division of labor leads to efficiency gains. First, dividing a larger task into smaller tasks allows each worker to gain skill in his designated task that he would not otherwise be able to attain. Additionally, reducing the number of tasks each worker must manage reduces transition times from one task to the

[1] Although Adam Smith popularized the notion of division of labor through his theory on the pin factory, he did not pioneer the notion. In 380 BC, Plato discussed in *The Republic* how the volume and quality of production could be improved through the division of labor (Silvermintz, 2010). This early discussion of the division of labor is not surprising given the intuitive nature of dividing tasks within a household and dividing occupations within a town. As technologies improved, the division of labor became more extensive. By the mid fifteenth century, The Venetian Arsenal was producing ships by using the river as an assembly line (Lane, 1992). Workers at each port were responsible for different parts of the ship that were added on as the ships moved down river. Sir William Petty (1992) documented similar innovations in the Dutch shipping industry in the seventeenth century.

next during which productivity is lost. Lastly, division of labor allows individuals to focus their full attention on a couple of simple tasks that increases the likelihood of technological innovation.

But pupils are not pins – and the production of human capital is far more complex than assembling automobiles. Whether specialization can increase productivity in schools is an important open question in the design of primary and secondary schooling. Indeed, there seems to be considerable disagreement across countries. Of the thirty-four OECD countries, only ten use specialized teachers in classrooms in elementary schools. Of the twenty-four countries that don't use specialized teachers Austria, Hungary, Norway, Portugal, Latvia, and Israel depart even further from teacher specialization. The average teacher in these countries stays with the same group of elementary school children for at least 3 years. This model of production is in stark contrast to how economists typically think about the division of labor, though consistent with the philosophical views of Marx (1844) and Thoreau (1854). If schools can increase the efficiency of human capital production by altering the allocation of teachers to subjects taught, simple policy changes might increase human capital at trivial costs.

Starting in the 2013-2014 school year, I conducted a randomized field experiment in fifty traditional public elementary schools in Houston, Texas, designed to test the potential productivity benefits of teacher specialization and shed light on what mechanisms drive the results. Treatment schools altered their schedules to have teachers specialize in a subset of the following subjects – math, science, social studies and reading – based on each teacher's strengths (assessed by the principal of each school). Schools then submitted specialization plans along with a written justification for each plan. Principals assigned teachers to subjects based on the principal's judgment of each teacher's comparative advantage. This judgment was based on either teacher value-added measures, classroom observations, or recommendations (for teachers new to the district or new to teaching).

In obtaining the optimal allocation of teachers to subjects, schools were constrained by how many teachers they had teaching a certain grade and language. The school district would not allow sorting teachers across schools, across grade-levels within a school, or across languages taught, because of the difficulties in extrapolating teacher effectiveness

across these categories.[2] With these constraints, there were 2-4 teachers available to teach a given grade and language group in over 80% of cases. Based on this availability, teams of teachers were designated within schools and grades. After reviewing school's departmentalization plans, we recommended further changes in teaching assignment for less than five percent of the cases and half of our recommendations were accepted; the teacher assignments in which principals did not accept our recommendation were all judgment calls for which we deferred. Control teachers continued the status quo.[3]

The results of the experiment are surprisingly *inconsistent* with the positive effects of specialization typically known to economists. In the first year of the experiment, the impact of attending an elementary school in which teachers were specialized was -0.073σ (0.036) in math and -0.067σ (0.030) in reading. In the second year of treatment, treatment effects were more precisely estimated zeros. Pooled across years, students in treatment elementary schools score 0.042σ (0.023) *lower* in math and 0.034σ (0.021) *lower* in reading, per year, relative to students in control elementary schools. The math score is statistically significant.

Students who might be particularly vulnerable – such as those enrolled in special education or those who are taught by inexperienced teachers– demonstrate particularly negative impacts of treatment. For special education students, the impact of treatment is –0.156σ (0.056) in math and -0.199σ (0.046) in reading. For non-special education students, the impact of treatment is -0.047σ (0.029) in math and –0.038σ (0.027) in reading. The p-value on the difference is 0.055 in math and 0.002 in reading. Students who were taught by younger teachers also demonstrated large negative treatment effects.

Beyond test scores, treatment schools, on average, are 0.016 more probable to have students exhibiting more problem behaviors and attending 0.35 fewer days of schools.

I argue that familiarity with student type explains a portion of the results. A benefit of specialization is that teachers are allowed to teach a subset of subjects in which they are (relatively) effective. A cost is that teachers in our experiments had, on an average, more

[2] Due to the large number of Spanish speaking students in Houston, there are bilingual classrooms, transitional bilingual classrooms, and English as a Second Language classrooms in elementary schools. Bilingual classrooms provide instruction primarily in Spanish in lower grades with increasing amounts of instruction in English added to instruction as the student advances to the upper grades. Transitional bilingual classrooms provide a bridge for limited English proficiency students to English-only instruction. The English as a Second Language program is offered to those students with a home language other than Spanish.

[3] If there were any departmentalized classes in control schools, they were kept as such.

total students in a day than in control schools – raising the costs of individually tailoring pedagogy. To better understand how the experiment altered teacher behaviors, a teacher survey was administered to glean information on lesson planning, teacher relationships with students, enjoyment of teaching, and teaching strategies. Teachers in treatment schools are significantly less likely to report providing tailored instruction for their students. All other survey outcomes on teaching strategy were statistically identical between treatment and control.

Taken together, the experiment highlights a potentially important tradeoff between the positive effects of specialization and the costs of tailoring pedagogical tools to fit student needs. I highlight this formally in the next section, which provides a brief review of the literature on the costs and benefits of specialization and combines the major hypotheses together in a simple equilibrium model. Section III provides details of the randomized field experiment and its implementation. Section IV describes the data, research design, and econometric model used in the analysis. Section V presents estimates of the effect of teacher specialization on student achievement and other outcomes. Section VI provides some discussion around how well the results of the experiment concord with the model. The final section concludes. There are two online appendices: Appendix A contains technical proofs; Appendix B describes how I construct my samples and defined key variables used in the analysis.

II. The 'Pupil' Factory

In this section, I review some of the major hypotheses about how teacher specialization may affect the production of human capital.

A. The Benefits of Teacher Specialization

Teacher specialization in schools may increase productivity for several reasons. First, if a teacher specializes in teaching a particular subject, there is more time to master subject specific content and pedagogy and more time to stay aware of advancements in the field. Second, specialization reduces the number of subjects a teacher is responsible for, allowing them to focus more energy on lesson planning and other subject-specific investments. Third, some argue that specialization increases teacher retention due to

reduced workload and reduced likelihood of teaching an unfamiliar subject.[4] Additionally, specialization offers a way to sort teachers by their comparative advantage and can increase – mechanically – average Teacher Value-Added (TVA) in each subject without having to make any staff changes. Finally, since specialization is the status quo in the upper grades, familiarizing students with it in elementary school may help ease the transition from elementary to middle school (Chan and Jarman 2004).

B. The Costs of Teacher Specialization

Becker and Murphy (1994) suggest that there are also potential costs to the division of labor; including coordination costs, principal agent problems between workers, and lack of economies of scale. Specialization occurs through the reorganization of existing staff. Teachers teach a larger number of students, but only teach a couple of subjects. Consequently, one potential cost is that each teacher has less time to get to know and understand any individual student (Anderson, 1962). This lack of information may increase the cost of tailoring pedagogy to fit student need. Additionally, specialization usually necessitates a student moving classrooms throughout the day. Frequent transitioning between classes may prevent teachers from having full information on a student's "state of the world" for that particular day. For instance, if pedagogical tool A is best used in state A and pedagogical tool B is best used in state B, having inferior information on the state of the world will yield inefficiencies in production.[5] Increased transition times between classes also decreases valuable instructional time (McGrath and Rust, 2002). Finally, teachers will have a harder time coordinating to ensure rules are enforced consistently and uniformly (Anderson, 1962). Behavior modification exercises such as assigning punishment based on a student's infractions for a day may be less effective when the teacher does not spend the full day with the student.

[4] Teacher retention was not significantly different between treatment and control schools. Table 5 displays the treatment effect on teacher retention in treatment versus control schools. The treatment effect for fraction of teachers retained between 2013-2014 and 2014-2015 is statistically insignificant.

[5] Relatedly, evidence suggests that there is a cost associated with care of patients across multiple physicians. For instance, a doctor giving continuous care to a patient will be more familiar with the patient's condition. After the doctor's shift, it may take time to update the new doctor on the patient's condition. Hence, some argue it is better for the entire care of a patient to be covered by a single physician rather than by specialists (Van Walraven et al. 2004). This intuition may be particularly important in other processes that also involve production of human capital where knowledge of an individual is an important input in production.

I cannot credibly identify the separate impact of each of these potential costs and benefits. Instead, this paper's goal is to produce credible estimates of the net impact of teacher specialization. The resulting "reduced form" will likely reflect a number of the potential channels highlighted above.

C. A Model

I now incorporate some insights from the literature into a simple model that is designed to better understand the experiment. As mentioned above, I cannot formally test between the various channels that teacher specialization may impact student achievement. Thus, I abstract away from all but the bare essentials in an effort to make the model crisp. The two key channels driving the tradeoff in the model are the benefits from specialization that accrue from sorting teachers based on their comparative advantage and costs of not tailoring pedagogy to student type. Adding the other channels into the model is trivial and not particularly illuminating.

The Basic Building Blocks

Let there be a large finite set, N, of agents referred to as "students" and one agent referred to as the "teacher." Nature moves first and assigns teaching knowledge H to the teacher, and a type, τ_j, to each student. I assume that student type is a pair (α_j, θ_j), where $\theta_j \in [\underline{\theta}, \overline{\theta}]$ represents innate ability and $\alpha_j \in [\underline{\alpha}, \overline{\alpha}]$ denotes a student idiosyncratic type, $j \in \{1,2, \ldots N\}$. Each student observes τ_j and chooses effort $e_j \in \mathbb{R}^+$.

The teacher observes his own teaching knowledge H, student's ability θ_j and student's effort level e_j. He does not observe students' idiosyncratic types α_j but instead, receives noisy signals $\{s_{j1}, s_{j2}, \ldots., s_{jT}\}$ about α_j for T time periods. Algebraically, signals are equal to the true idiosyncratic types plus some normal noise: $s_{jt} = \alpha_j + \varepsilon_{jt}$, where $\varepsilon_{jt} \sim N(0, \sigma_\varepsilon^2)$.

After receiving $N \times T$ signals about α_j's from N students for T time periods, the teacher "sets a dial" $x \in \mathbb{R}$. This assumption is motivated by the model described in Jovanovic and Rousseau (2001). One can think of setting dial x, in this context, as the

teacher choosing his teaching pedagogy to maximize total student achievement.

Payoffs

We assume that student achievement is related to observable and unobservable parameters in the manner described in Jovanovic and Rousseau (2001). Let Y_j denote student achievement of pupil *j*: $Y_j = f(e_j, \theta_j, H) - (\alpha_j - x)^2$, where f is smooth, continuous, and increasing in its arguments. Thus, higher student achievement can be achieved by either increasing student effort e_j, ability θ_j, or teacher's knowledge H, and by decreasing the absolute distance between the teacher's choice of pedagogy x and the student's idiosyncratic type α_j. Therefore, the payoff that the teacher receives by setting a dial $x \in \mathbb{R}$ is equal to the total achievement of N students: $\sum_{j=1}^{N} Y_j = f(e_j, \theta_j, H) - (\alpha_j - x)^2$.[6]

Student payoffs depend on how much effort is exerted and the costs and benefit of that effort choice. In symbols: $f(e_j, \theta_j, H) - (\alpha_j - x)^2 - k(e_j)$, where $k(e_j)$ denotes costs of effort. I assume that costs of effort are increasing and convex: $\frac{\delta k(e_j)}{\delta e_j} > 0 \; and \; \frac{\delta^2 k(e_j)}{\delta {e_j}^2} > 0$.

Strategies

The teacher's strategy is to choose a teaching pedagogy $x: \mathbb{R}^{NXT} \rightarrow \mathbb{R}$ after observing $N \times T$ signals about α_j's from N students for T time periods. A student's strategy is a mapping from their innate ability to an effort choice: $e: [\underline{\theta}, \overline{\theta}] \times [\underline{\alpha}, \overline{\alpha}] \rightarrow \mathbb{R}^+$.

Expected Payoffs

The teacher maximizes expected student achievement from his class after observing signals about students' α_j's. Let S_j denote a $1 \times$T vector of signals for student *j*. Total expected student achievement conditional on observing a stream of signals S_j is given by:

$$\sum_{j=1}^{N} E(Y_j | S_j) = \sum_{j=1}^{N} f(e_j, \theta_j, H) - E\left((\alpha_j - x)^2 \mid S_j\right). \quad (1)$$

[6] In a previous version of the model, I allowed student effort to depend on the dial set, x, by writing the payoff function as: $\sum_{j=1}^{N} Y_j = f(e_j, \theta_j)\left(H - (\alpha_j - x)^2\right)$. This model is significantly more complicated but yields the same qualitative results.

If the signal vector provided the teacher full information on students' idiosyncratic types, it is straightforward to demonstrate that, when maximizing student achievement, the teacher sets the dial equal to the weighted average of α_j's: $\sum_{j=1}^{N} \frac{\alpha_j}{N}$.

However, by assumption, the teacher does not receive full information on students' idiosyncratic types. He has some prior beliefs about types and updates his beliefs according to Bayes rule. To illustrate, assume $t = 1$ and let the teacher's prior about any student's idiosyncratic type be given by $\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$. At the end of the period, the teacher receives s_{j1} for each student and updates his prior on α_j. Since the normal learning model is generally easier to think about in terms of precision, let $h = \frac{1}{\sigma^2}$ denote a measure of how tight a distribution is.

Manipulating notation, the signal and the prior can be written as: $s_{jt} \sim N\left(\alpha_j, \frac{1}{h_\varepsilon}\right)$ and $\alpha_j \sim N\left(\mu_\alpha, \frac{1}{h_\alpha}\right)$, respectively. The posterior belief about α_j after receiving signal s_{j1} is given by $posterior_{t=1}$: $\alpha_j | s_{j1} \sim N\left(\frac{h_\alpha \mu_\alpha + h_\varepsilon s_{j1}}{h_\alpha + h_\varepsilon}, \frac{1}{h_\alpha + h_\varepsilon}\right)$. Extending to $t = T$ and deploying a bit of algebra, one can rewrite this as:

$$E\left(\alpha_j \middle| s_{j1,} s_{j2,} \ldots s_{jT}\right) = \frac{h_\alpha \mu_\alpha + h_\varepsilon \sum_{t=1}^{T} s_{jt}}{h_\alpha + T h_\varepsilon} \quad (2)$$

Equilibrium

An equilibrium is a pair of strategies, x^* and e_j^*, for all *j*, such that each is a best response to the other. Assuming risk neutrality, the teacher will choose x to maximize total expected student achievement. Using equations (1 and 2), the optimal dial is $x^*(S_j) = \frac{\sum_{j=1}^{N} E(\alpha_j | S_j)}{N}$, where $E\left(\alpha_j \middle| S_j\right) = \frac{h_\alpha \mu_\alpha + h_\varepsilon \sum_{t=1}^{T} s_{jt}}{h_\alpha + T h_\varepsilon}$. Equilibrium student effort can be written as the $e^* \rightarrow \frac{\delta f}{\delta e_j} - \frac{\delta k}{\delta e_j} = 0$.

Teacher Specialization

In the current model, teacher specialization is akin to receiving fewer signals about students' idiosyncratic types. In traditional elementary classrooms, teachers are with the same set of students all day. Conversely, when teachers specialize, they teach a subset of subjects

(half, say) and teach significantly more students (double, say). For simplicity and transparency, I assume that without teacher specialization, $T \to \infty$ and with teacher specialization, $T \to 0$.[7]

Proposition 1: *With teacher specialization, total student achievement increases if teacher's knowledge, H, increases such that*:

$$\sum_{j=1}^{N} f(e_j, \theta_j, H_s) - \sum_{j=1}^{N} f(e_j, \theta_j, H_{ns}) > N\left(\mu_\alpha - \sum_{j=1}^{N} \frac{\alpha_j}{N}\right)^2 \quad (3)$$

Proof – See Appendix A.

Proposition 1 provides a formal description of the costs and benefits of teacher specialization. In words, the proposition highlights that student achievement will increase under specialization whenever the human capital benefit of sorting teachers based on comparative advantage – $\sum_{j=1}^{N} f(e_j, \theta_j, H_s) - \sum_{j=1}^{N} f(e_j, \theta_j, H_{ns})$ is larger than the cost of inefficient "dial setting" -- $N\left(\mu_\alpha - \sum_{j=1}^{N} \frac{\alpha_j}{N}\right)^2$. This provides the essence of the problem.[8]

Other potential costs of specialization – such as less time with teachers due to frequent classroom transitions – can be added without changing the basic economics. A similar argument applies to the benefit side. For instance, a potentially important benefit of specialization is that teachers have more time to master pedagogical tools specific to their subjects. I have assumed that teacher capacity is fixed. In a fuller model, one might allow the law of motion of teacher ability to be affected by the number of classes they teach.

III. Background and Field Experiment Details

Houston Independent School District (HISD) is the seventh largest school district in America with more than 200,000 students in almost 300 schools. Eighty-eight percent of HISD students are black or Hispanic. Approximately 80 percent of all students are eligible

[7] These limiting cases are a matter of mathematical convenience. The results also hold for any $T > 0$ if one assumes that teachers who are specialized have a signal vector that is first order stochastically dominated by the signal vector received by non-specialized teachers.

[8] One might argue that the formulation of the tradeoff is more about different types of specialization – there might be specialization in the traditional sense of comparative advantage and specialization in the task of getting to know students in a more nuanced manner. Since this is more about semantics than substance, I chose to articulate the decision in the starkest terms.

for free or reduced-price lunch and roughly 30 percent of students have limited English proficiency.

To begin the field experiment, we followed standard protocol. First, we garnered support from the district superintendent and other key district personnel. The district then provided a list of 62 schools that were eligible for randomization into the teacher specialization experiment.[9] I removed twelve of these schools because either they were part of another experiment or because their particular school model was antithetical to the notion of teacher specialization (e.g. Montessori).[10] Our final experimental sample consists of fifty schools – twenty-five treatment and twenty-five control – that were randomly allocated vis-à-vis a matched-pair procedure (details to follow).

After treatment and control schools were chosen, treatment schools were alerted that they would alter their schedules to have teachers specialize in a subset of the following subjects – math, science, social studies and reading – based on each teacher's strengths. Schools then sent in specialization plans along with a written justification for each plan. Principals assigned teachers to subjects based on the principal's perception of each teacher's comparative advantage. This perception was based on either TVA measures, classroom observations, or recommendations (for teachers new to the district or new to teaching).

Schools were constrained as to how many teachers they had teaching a certain grade and language since teachers were prohibited from switching between these categories. Given these grade-level and language constraints, there were 2-4 teachers available to teach a given grade and language group in over 80% of cases. Based on this availability, teams of teachers were designated within schools and grades. Teachers were not permitted to teach both math and reading. In the modal case of a two teacher team, one teacher taught math and science and one teacher taught reading and social studies. Otherwise, one teacher taught reading, one teacher taught math, and the teachers shared teaching duties for social studies and science. Some teacher teams had three teachers where one taught math, one taught reading and the third taught science and social studies. Students had different teachers for different subjects, but stayed with the same group of classmates for all subjects.

[9] When choosing a list of experimental schools, the district, besides allowing for schools with minority and low achieving students, focused on schools that had the capacity to sort teachers to teach specialized subjects

[10] Montessori education encourages teachers to teach for long hours fostering the development of environments that develop a child's natural psychological, physical and social development (http://amshq.org)

After reviewing schools' departmentalization plans, I recommended further changes in teaching assignment for 25 out of 520 teachers. I made recommendations for changes only in cases where the principal's decision seemed to contradict Houston's calculated TVA for the 2011-2012 school year or author-calculated TVA for 2012-2013 school year. Schools then sent updated departmentalization plans and 14 of our recommended changes were agreed upon by the school. In the remaining eleven cases, the principals indicated their choices and arguments justifying their decisions. For instance, we recommended that a 3rd grade teacher might be better suited to reading than math in a particular elementary school but the school decided to keep original assignments stating that the teacher was better suited to math based on summer school observations.[11]

Table 1 describes differences between treatment and control elementary schools and all other elementary schools in HISD across a set of covariates gleaned from administrative data. The descriptive differences between participating (treatment and control) and non-participating schools is consistent with the fact that the leadership of HISD preferred elementary schools that were predominantly minority and low-achieving to enter the experimental sample. Students in experimental schools are less likely to be white, more likely to be black, less likely to be Asian, more likely to be economically disadvantaged, more likely to be in a special education program, less likely to be gifted, and have lower pre-treatment test scores in math and reading. Thus, the results estimated are likely more applicable to urban schools with high concentrations of minority students.

IV. Data, Research Design, and Econometric Framework

Data

We use administrative data provided by the Houston Independent School District (HISD). The main HISD data file contains student-level administrative data on approximately 200,000 students across the Houston metropolitan area in a given year. The data includes information on student race, gender, free and reduced-price lunch status, behavior, and attendance for all students; state math and reading test scores for students in third through fifth grades; and Stanford 10 subject scores in math and reading for elementary school students. Behavior data records student behavioral incidents resulting in a serious disciplinary

[11] Based on survey responses, the average fraction of specialized teachers in control schools is 54.2%. The corresponding fraction in treatment schools is 89.9%.

action such as a suspension or an expulsion. I have HISD data spanning the 2010-2011 to 2014-2015 school years. I also collected data from a survey administered to teachers at the end of the 2013-2014 school year. 418 (80 % response rate) treatment teachers and 343 (70% response rate) control teachers completed the survey. See Online Appendix A for details on the outcomes used from the survey.

The state math and reading tests, developed by the Texas Education Agency (TEA), are statewide high-stakes exams conducted in the spring for students in third through eleventh grade.[12] Students in fifth grade must score proficient or above on both tests to advance to the next grade. Because of this, students in the fifth grade who do not pass the tests are allowed to retake it approximately one month after the first administration. We use a student's first score unless it is missing.[13]

All public school students are required to take the math and reading tests unless they are medically excused or have a severe disability. Students with moderate disabilities or limited English proficiency must take both tests, but may be granted special accommodations (additional time, translation services, alternative assessments, and so on) if they meet certain requirements set by the Texas Education Agency. In our analysis, the test scores are normalized (across the school district) to have a mean of zero and a standard deviation of one for each grade and year.[14]

We use a parsimonious set of controls to help correct for pre-treatment differences between students in treatment and control schools. The most important controls are reading and math achievement test scores from the three years *prior to the start of the experiment*, which we include in all regressions (unless otherwise noted), and are also referred to throughout the text as "baseline test scores". We also include one indicator variable for each baseline test score that takes on the value of one if that test score is a Spanish version test and zero

[12] Sample tests can be found at http://www.tea.state.tx.us/student.assessment/released-tests/ .
[13] Using their retake scores, when the retake is higher than their first score, does not significantly alter the results. Results available from the author upon request.
[14] Among students who take a state math or reading test, several different test versions are administered to accommodate specific needs. These tests are designed for students receiving special education services who would not be able to meet proficiency on a similar test as their peers. STAAR-- L is a linguistically accommodated version of the state mathematics, science and social studies test that provides more linguistic accommodations than the Spanish versions of these tests. According to TEA, STAAR--Modified and STAAR--L are not comparable to the standard version of the test and thus, we did not use them for our main analysis. We did, however, investigate whether treatment influenced whether or not a student takes a standard or non-standard test (see Appendix Table 1).

otherwise. Baseline scores are STAAR test scores for students in grades three through five in the baseline year and Stanford 10 for students in grade K-2 in the baseline year.

Other individual-level controls include gender; a mutually exclusive and collectively exhaustive set of race indicator variables; and indicators for whether a student is eligible for free or reduced-price lunch or other forms of federal assistance, whether a student receives accommodations for limited English proficiency, whether a student receives special education accommodations, or whether a student is enrolled in the district's gifted and talented program.[15]

To supplement HISD's administrative data, a survey was administered to all teachers in both treatment and control at the end of the 2013-2014 school year. The data from the survey includes questions about lesson planning, relationship with students and interaction with parents and guardians of students. Teachers were given a $20 Amazon.com gift card for completing the survey and principals were informed that they would also receive a $40 Amazon.com gift card if they were able to get teacher participation above 80% at their campus. Approximately seventy percent of control teachers completed the survey while the corresponding fraction for treatment teachers was eighty percent.

Research Design

To partition the set of interested schools into treatment and control, we used a matched-pair randomization procedure. Recall, fifty schools entered our experimental sample from which we constructed twenty-five matched pairs. Following the recommendations in (Abadie and Imbens, 2011), control and treatment groups were balanced on a variable that was correlated with the outcomes of interest – past standardized test scores. First, the full set of fifty schools were ranked by the sum of their mean reading and math test scores in the previous two years. Then, we designated every two schools from

[15] A student is income-eligible for free lunch if her family income is below 130 percent of the federal poverty guidelines, or categorically eligible if (1) the student's household receives assistance under the Food Stamp Program, the Food Distribution Program on Indian Reservations (FDPIR), or the Temporary Assistance for Needy Families Program (TANF); (2) the student was enrolled in Head Start on the basis of meeting that program's low-income criteria; (3) the student is homeless; (4) the student is a migrant child; or (5) the student is identified by the local education liaison as a runaway child receiving assistance from a program under the Runaway and Homeless Youth Act. HISD Special Education Services and the HISD Language Proficiency Assessment Committee determine special education and limited English proficiency status respectively.

this ordered list as a "matched pair" and randomly selected one member of the matched pair into the treatment group and one into the control group.

Columns (1) and (2) of Table 2 display descriptive statistics on individual student characteristics of all HISD students in third through fifth grade. The first column provides the mean for each variable for control school students. The second column provides the difference between the treatment and the control group, which we estimate by regressing the variable on a treatment indicator and matched pair fixed effects. Of the 14 student-level variables, only 1 is statistically significant at the 5% level. 5.6 percent of control students record a behavioral incident while 8.3 percent of treatment students record being involved in a behavioral incident.[16]

Econometrics

To estimate the causal impact of our treatment on outcomes, we estimate both intent-to-treat (ITT) effects and Local Average Treatment Effects (LATEs). Let Z_i be an indicator for assignment to treatment, let X_i denote a vector of baseline variables (consisting of the demographic variables in Table 2) measured at the individual level, let $f(\cdot)$ represent a polynomial including 3 years of individual test scores in both math and reading prior to the start of treatment and their squares. *All of these variables are measured pre-treatment.* Moreover, let γ_g denote a grade-level fixed effect and Ψ_m a matched-pair fixed effect.

The Intent-to-Treat (ITT) effect, τ_{ITT}, using the twenty-five treatment and twenty-five control schools in our experimental sample can be estimated with the following equation:

$$(4) \qquad Y_{i,\,m,\,g,\,yr} = a + \tau_{ITT} \cdot Z_i + f(Y_{i,\,TR\text{-}1},\, Y_{i,\,TR\text{-}2},\, Y_{i,\,TR\text{-}3}) + \beta X_i + \gamma_g + \Psi_m + \eta_{yr} + \varepsilon_{i,\,m,\,g,\,yr}$$

where *TR* represents the treatment year.

Equation (4) identifies the impact of being *offered a chance* to attend a treatment school, τ_{ITT}, where students in the matched-pair schools correspond to the counterfactual state that would have occurred for the students in treatment schools had their school not been randomly selected. I focus on a fixed population of students. A student is considered treated (resp.

[16] See Online Appendix A for details on how each variable was constructed.

control) if they were in a treatment (resp. control) school in the pre-treatment year and not in an exit grade (e.g. 5th grade). All student mobility after treatment assignment is ignored. Note that because Equation (4) is estimated on first through fifth graders and treatment assignment was determined in the pre-treatment year, students selecting into treatment is not a concern.

Yet, in any experimental analysis, a potential threat to validity is selection out of sample. For instance, if schools that implement teacher specialization are more likely to have low (resp. high) performing students exit the sample, then our estimates will be biased upwards (resp. downwards) – even under random assignment. We find that 9.27% of treatment student observations are missing a test score relative to 10.92% of control students, a difference of 1.65%. Thus, despite attrition rates being around 10.06%, the difference in attrition between treatment and control is sufficiently small that Lee (2009) bounds on treatment effects remain qualitatively the same – and quantitatively similar – as the ITT treatment effects. This issue is addressed in more detail in the following section.

Under several assumptions (e.g. that treatment assignment is random, control schools are not allowed to participate in the program and treatment assignment only affects outcomes through program participation), we can also estimate the causal impact of *attending* a treatment school. This parameter, commonly known as the Local Average Treatment Effect (LATE), measures the average effect of attending a treatment school on students who attend as a result of their school being randomly selected. We estimate two different LATE parameters through two-stage least squares regressions, using random assignment as an instrumental variable for the first stage regression. The first LATE parameter uses an indicator variable, $EVER$ which is equal to one if a student attended a treatment school for at least one day. More specifically, in the 2014 specification, $EVER$ is equal to one if a student attended a treatment school for at least one day in the 2013-2014 school year and zero otherwise and uses test scores from 2014 as an outcome. In the 2015 specification, $EVER$ is equal to one if a student attended a treatment school for at least one day in 2013-2014 or 2014-2015 and zero otherwise and uses test scores from 2015 as an outcome. In the pooled specification, $EVER$ is equal to one if a student attended a treatment school for at least one day in 2013-2014 or 2014-2015 and zero otherwise and uses test scores from both 2014 and 2015 as an outcome. The second stage equation for the two-stage least squares estimate therefore takes the form:

$$(5) \qquad Y_{i,\,m,\,g,\,yr} = a + \Omega EVER_{i,\,m,\,g,\,yr} + f(Y_{i,\,TR\text{-}1}\,,\, Y_{i,\,TR\text{-}2}\,,\, Y_{i,\,TR\text{-}3}) + \beta X_i + \gamma_g + \eta_{yr} + \Psi_m + \varepsilon_{i,\,m,\,g,\,yr}$$

and the first stage equation is:

$$(6) \qquad EVER_{i,\,m,\,g,\,yr} = a + \lambda Z_i + f(Y_{i,\,TR\text{-}1},\, Y_{i,\,TR\text{-}2},\, Y_{i,\,TR\text{-}3}) + \beta X_i + \gamma_g + \eta_{yr} + \Psi_m + \varepsilon_{i,\,m,\,g,\,yr}$$

where all other variables are defined in the same way as in Equation (1). When Equation (5) is estimated for one year only, Ω (referred to as 2SLS (Ever) in tables) provides the cumulative treatment effect in that year. When Equation (5) is estimated across multiple years, as in the pooled estimates, Ω provides the weighted average of the cumulative effects of attending a treatment school.

Our second LATE parameter is estimated through a two-stage least squares regression of student achievement on the intensity of treatment. More precisely, we define *TREATED* as the number of years a student is present at a treatment school. The second stage equation for the two-stage least squares estimate therefore takes the form:

$$(7)\ Y_{i,\,m,\,g,\,yr} = a + \delta TREATED_{i,\,m,\,g,\,yr} + f(Y_{i,\,TR\text{-}1},\, Y_{i,\,TR\text{-}2},\, Y_{i,\,TR\text{-}3}) + \beta X_i + \gamma_g + \eta_{yr} + \Psi_m + \varepsilon_{i,\,m,\,g,\,yr}$$

and the first stage equation is:

$$(8)\ TREATED_{i,\,m,\,g,\,yr} = a + \lambda \cdot Z_i + f(Y_{i,\,TR\text{-}1},\, Y_{i,\,TR\text{-}2},\, Y_{i,\,TR\text{-}3}) + \beta X_i + \gamma_g + \eta_{yr} + \Psi_m + \varepsilon_{i,\,m,\,g,\,yr}$$

The first stage equation is equivalent to Equation (6), but with *TREATED* as the dependent variable. In the 2014 specification, *TREATED* ranges from zero to one and uses test scores from 2014 as an outcome. In the 2015 specification, *TREATED* ranges from zero to two and uses test scores from 2015 as an outcome. In the pooled specification, *TREATED* ranges from zero to two and uses test scores from both 2014 and 2015 as an outcome. Therefore, δ provides the average yearly effect of participating in our experiment.

V. Teacher Specialization and the Production of Human Capital in Schools

Table 3 presents a series of estimates of the impact of teacher specialization on reading and math state test scores. The rows provide estimates for different outcomes and each set of columns coincides with a different empirical model that is being estimated. Test

scores are normalized to have mean 0 and standard deviation 1 across the entire school district by grade and year, so treatment effects are presented in standard deviation units. Standard errors, clustered at the school level, are in the parentheses below each estimate along with the number of observations. All regressions include grade, year and matched-pair fixed effects.

Columns (1) and (2) present ITT estimates of the impact of teacher specialization on math and reading achievement state test scores for the 2013-2014 and 2014-2015 school years, respectively. In the first year of the experiment, the impact of being offered the chance to attend an elementary school in which teachers were specialized was -0.062σ (0.031) in math and -0.057σ (0.026) in reading. In the second year of treatment, treatment effects were smaller and measured with a bit more noise which rendered them statistically insignificant. Pooling across years, which is shown in column (3), students in treatment elementary schools score 0.051σ (0.028) *lower* in math and 0.041σ (0.027) *lower* in reading relative to students in control elementary schools. The math score is statistically significant.

Columns (4) and (5) present LATE estimates for the cumulative effect of actually attending a treatment school for at least one day in one of the school years. Column (6) contains the pooled estimate. The average cumulative effect of attending a treatment school for at least one day in any of the school years is -0.058σ (0.032) in math and -0.046σ (0.030) in reading. Columns (7) through (9) present yearly LATE estimates which capture the effect of actually attending any treatment school. Thus, to calculate the total effect of the intervention one multiplies the estimates by two. The impact of teacher specialization is –0.042σ (0.023) in math and -0.034σ (0.021) in reading, per year. Thus, at the end of the two-year experiment encompassing 18 school months, students with specialized teachers were approximately one month behind students with non-specialized teachers – implying that specialization reduces production efficiency by 6 percent.

These results are surprisingly *inconsistent* with the positive effects of division of labor typically known to economists though, as Proposition 1 illustrates, might be consistent with a model in which specialization results in inefficient dial setting – though other mechanisms are possible.

Another, perhaps even more transparent, way to look at the data is to graph the distribution of treatment effects for each matched pair-grade cell, which is depicted in Figure

1. We control for demographic observables and baseline test scores by estimating equation (4) for each matched pair and grade. We then collect the treatment coefficients from this equation and plot a kernel density curve for them. The results echo those found in Table 3. In math, 38 out of 75 match pair by grade level cells have negative results. In reading, 35 out of 75 match pair by grade level cells have negative effects.

Attendance and Behavioral Incidence

Consistent with the impact on test scores, there is a positive effect on student suspensions and a negative effect of treatment on attendance. Specifically, treatment schools, on average, have a 0.016 (0.001) higher probability of student suspension due to poor behavior in the treatment year and 0.003 (0.000) lesser number of years in attendance.

Heterogeneous Treatment Effects

Table 4 explores the sensitivity of the estimated treatment effects across pre-determined subsamples of the data. The negative effects of teacher specialization are remarkably robust, though there is some evidence that students who are more likely to need individual attention – e.g. students with special needs – do particularly poorly when teachers are specialized. The coefficient on treatment for students with special needs is -0.156σ (0.056) and the effect for students without special needs is -0.047σ (0.029) in math; the respective estimates in reading are -0.199σ (0.046) and -0.038σ (0.027). Both differences are statistically significant.

The results become more interesting when the data are partitioned by pre-treatment teacher age at the time of intervention. Teachers age at the time of intervention was culled from the district's administrative records. Consistent with the findings above, treatment effects are more negative and pronounced for younger teachers – those who one might consider more "at-risk." Dividing the sample of teachers into terciles – based on their age – we find that the treatment effect on the youngest tercile of teachers is -0.174σ (0.041) in math and -0.208σ (0.052) in reading. In comparison, the treatment effect on the oldest tercile of teachers is 0.025σ (0.063) in math and -0.077σ (0.078) in reading. The p-value for the treatment coefficient being different for all three terciles is 0.000 in math and 0.030 in reading.

VI. Robustness Checks

In this subsection we explore the robustness of our results under two potential threats to our interpretation of the data.

Attrition and Bounding

A concern for estimation is that we only include students for which we have post-treatment test scores. If students in treatment schools and students in control schools have different rates of selection into this sample, our results may be biased. Appendix Table 1 compares the rates of attrition of students in treatment schools and students in control schools. The first panel uses whether or not a student has a missing math score as an outcome. The numbers reported in the columns (2), (4) and (6) are the coefficients on the treatment indicator. The second panel has whether or not a student has a missing reading score as an outcome. To see whether attrition affects our estimates, we compute Lee (2009) bounds in Appendix Table 2, which calculates conservative bounds on the true treatment effects under the assumption that attrition is driven by the same forces in treatment and control, but that there are differential attrition rates in the two samples. Under the Lee method, children are selectively dropped from either the treatment or control group to equalize response rates. This is accomplished by regressing the outcome variable on baseline controls and treatment status, and storing the residuals. When the probability of missing an outcome is higher for the control group, then treatment children with the *highest* residuals are dropped. When the probability of missing an outcome is higher for the treatment group, then control children with the *lowest* residuals are dropped. In our case, however, because the attrition rates are quite similar between treatment and control, qualitatively the treatment effects remain unchanged.

Alternative Specifications

In our main analysis we use matched-pair fixed effects and clustered standard errors as a way of obtaining consistent standard errors. Yet, this may not correct for school-level heterogeneity. This heterogeneity is uncorrelated with treatment due to random assignment, but could affect inference (Moulton 1986, 1990). Table 5 presents estimates after running

(population-weighted) school level regressions of the impact on test scores in the treatment year.

The pooled ITT effect on math scores is -0.083σ (0.002) and on reading scores is -0.071σ (0.002). If anything, these school-level regressions are more negative than the estimates at the individual level. Similarly, the impact of treatment on behavioral incidence is 0.016 (0.001) and -0.003 (0.000) on attendance.

Alternative 'Low Stakes' Test Scores

Jacob (2005) demonstrates that the introduction of accountability programs increases high-stakes test scores without increasing scores on low-stakes tests, most likely through increases in test-specific skills and student effort. It is important better understand whether the results presented above are being driven by actual losses in general knowledge or whether specialized teachers are simply more inefficient at high-stakes test preparation.

To provide some evidence on this question, I present data from the Stanford 10. Houston is one of a handful of cities that voluntary administers a nationally normed test for which teachers and principals are not held accountable – decreasing the incentive to teach to the test or engage in other forms of manipulation. The math and reading tests are aligned with standards set by the National Council of Teachers of Mathematics and the National Council of Teachers of Reading, respectively.[17]

Table 6 presents estimates of impact of teacher specialization on Stanford 10 math and reading scores. As in the state test results, the impact of teacher specialization is, if anything, negative. Panel A displays results for grades three through five. This sample matches the one in the main specifications as state tests are only administered in Houston in grades three through eight. Columns (1) and (2) present ITT estimates of the impact of teacher specialization on math and reading achievement state test scores for the 2013-2014 and 2014-2015 school years, respectively. In the first year of the experiment, the impact of being offered the chance to attend an elementary school in which teachers were specialized

[17]Math tests include content testing number sense, pattern recognition, algebra, geometry, and probability and statistics, depending on the grade level. Reading tests include age-appropriate questions measuring reading ability, vocabulary, and comprehension. More information can be found at http://www.pearsonassessments.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=SAT10C.

was -0.061σ (0.027) in math and -0.061σ (0.027) in reading. These estimates are nearly identical to those from the main specifications displayed in Table 3.

In the second year of treatment, treatment effects were smaller and measured with a bit more noise which rendered the math effect statistically insignificant. Pooling across years, which is shown in column (3), students in treatment elementary schools score 0.040σ (0.027) *lower* in math and 0.056σ (0.023) *lower* in reading relative to students in control elementary schools. The reading score is statistically significant. The results from first and second grade, displayed in panel B, are qualitatively similar but have even larger standard errors.

Permutation Tests

We have run several regressions with various outcomes in differing subsamples to measure treatment effects. The problem of multiplicity can lead one to incorrectly reject some null hypothesis purely by chance. To correct for this, we conduct a nonparametric permutation test as in Rosenbaum (1988). We re-randomized the sample 10,000 times between matched pairs at the school-level, like the original randomization. We re-ran the ITT regressions with the new, synthetic treatment assignments and recorded the new treatment effects. Appendix Figures 1 plot the actual observed ITT treatment effect against the distribution of simulated treatment effects for various outcomes. The key results with strong statistical significance – for instance, the negative effect on reading and math scores in the first year of treatment – are easily visualizable in Appendix Figures 1. Together, these results confirm the basic facts described throughout; teacher specialization, if anything, lowers student achievement.

VII. Interpreting the Data Through the Lens of the Dial-Setting Model

The experiment designed and evaluated in the previous sections generated a set of new facts. Sorting teachers in a way that allows them to teach a subset of subjects of relative strength has, if anything, negative impacts on test scores, negative impacts on attendance, and increases suspensions due to ill-advised behavior. Moreover, these impacts seem particularly stark for students with special needs and students taught by younger teachers.

Recall, Proposition 1 describes the conditions under which teacher specialization may lead to higher academic achievement. The key inequality is: (A) the increase in teacher

knowledge due sorting on comparative advantage versus (B) suboptimal pedagogy due to inefficient dial-setting. On one side of the ledger, (A), specialization should lead teachers with weakly higher TVA scores to provide better instruction to students. On the other side, (B), because teachers have less time, and hence, fewer interactions with their students they may "set the dial" sub-optimally leading to less effective instructional strategies.

The increase in student achievement caused due to teacher sorting on comparative advantage can be indirectly computed. Panel C in table 4 displays treatment effects for subgroups divided on the basis of how different a teacher's TVA is in the subject that he is sorted to teach versus the average TVA across all subjects the teacher used to teach before. We create terciles based on this difference and conduct ITT regressions on the pooled sample. As table 4 shows, teachers who are in the first tercile (or, who stand to gain the least from sorting) have the largest *decreases* in student achievement. For math, the treatment effect in this tercile is -0.185σ (0.031) while for reading, the treatment effect is -0.104σ (0.047). The corresponding treatment effects for the third tercile, or for teachers who stand to gain the most from sorting, is 0.008σ (0.044) in math and 0.035σ (0.034) in reading. Treatment effects across the three terciles are significantly different from each other at the 5% level.

Unfortunately, it is exceedingly difficult to test whether or not teachers correctly "set the dial." The survey evidence collected from treatment and control teachers provides an indirect way of assessing this portion of the theory. Survey data was collected at the end of 2013-2014 school year and was designed specifically to gather information on teaching strategies and interactions between teachers and their students.

Appendix table 3 reports treatment effects from our least-squares ITT specification (in column 2) and a Logistic regression (in Column 3) on teaching pedagogies: whether the teacher had personal relationships which each of his students, if the teacher feels that he gives students' individual attention, if rules are consistently enforced in the school, and if the teacher is enthusiastic about teaching a subject. For each of these outcomes, a variable is coded as 1 if teacher agrees with the statement to any extent and 0 otherwise. I also present impacts for the percentage of time spent on lesson differentiation for treatment versus control teachers.

There is some suggestive evidence that inefficient dial-setting may explain a portion of the results. Treatment teachers are 0.02 (0.03) less likely to report they "know" their students (control mean = 81.6%) and 0.041 (.02) less likely to report providing them with

individual attention (control mean = 62.5%). Only the latter is statistically significant. In contrast, there was no effect of treatment on whether rules are consistently enforced in school, teacher's reported enthusiasm for teaching or how much they attempted to differentiate their lessons.

These data are broadly consistent with the model developed in Section II, or any model in which having less time and attention to devote to each student is a cost of teacher specialization.

An important caveat to the above survey results is that there is a ten percentage point difference between treatment and control in response rates to the survey. Thus, any standard bounding procedure that takes this differential response rate into account will contain estimate intervals that are too large to be informative (See Appendix table 4).

VIII. Conclusion

Division of labor is a basic economic concept – the power of which, to date, has not been quantified vis-à-vis the production of human capital. In simple production processes – such as pins – there can be large positive gains from specialization. In schools however, having teachers specialize may increase the quality of human capital available to teach students through sorting, but may lead to inefficient pedagogical choices.

Empirically, I find that teacher specialization, if anything, decreases student achievement, decreases student attendance, and increases student behavioral problems. This result is consistent with the dial-setting model if teachers received fewer signals about their students' types after being departmentalized and the change in teacher value-added due to sorting was not large enough. I provide some suggestive evidence for this, though other mechanisms are possible.

These results provide a cautionary tale about the potential productivity benefits of the division of labor when applied to human capital development.

References

Abadie, A., & Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*.

Anderson, R. C. (1962). The case for teacher specialization in the elementary school. *The Elementary School Journal*, 253-260.

Becker, G. S., & Murphy, K. M. (1994). The division of labor, coordination costs, and knowledge. In *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education (3rd Edition)* (pp. 299-322). The University of Chicago Press.

Chan, T. C., & Jarman, D. (2004). Departmentalize Elementary Schools. *Principal*, *84*(1), 70-72.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of public Economics*, *89*(5), 761-796.

Jovanovic, Boyan, and Peter L. Rousseau. (2001). Why Wait? A Century of Life before IPO. *American Economic Review*, 91(2): 336-341.

Lane, F. C. (1992). *Venetian ships and shipbuilders of the Renaissance*. John Hopkins University Press.

Lee, David S. (2009). "Training, wages, and sample selection: Estimating sharp bounds on treatment effects". *The Review of Economic Studies*, *76*(3), 1071-1102.

Marx, K. (2012). *Economic and philosophic manuscripts of 1844*. Courier Corporation.

McCalley, Bruce W. (1989). The Model T Ford Encyclopedia, 1909-1927: A comprehensive guide to the volution and changes of the major componenets of the Model T Ford. *Model T Ford Club of America.* 1989

McGrath, C. J., & Rust, J. O. (2002). Academic achievement and between-class transition time for self-contained and departmental upper-elementary classes. *Journal of Instructional Psychology*, *29*(1), 40.

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of econometrics*, *32*(3), 385-397.

Moulton, B. R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The Review of Economics and Statistics*, 334-338.

Petty, W. (1992). *Political arithmetick, or a discourse concerning the extent and value of lands, people [and] buildings*. R. Clavel.

Rosenbaum, Paul R. (1988). Permutation tests for matched pairs with adjustments for covariates. *Applied Statistics* (1988): 401-411.

Silvermintz, D. (2010). Plato's Supposed Defense of the Division of Labor: A Reexamination of the Role of Job Specialization in the Republic. *History of Political Economy*, *42*(4), 747-772.

Smith, A. (1937). *An Enquiry into the wealth of nations [1776]*. Strahan and Cadell, London.

Thoreau, H. D. (1854). *Walden; or, Life in the Woods.*Boston, MA: Ticknor and Fields.

Van Walraven, C., Mamdani, M., Fang, J., & Austin, P. C. (2004). Continuity of care and patient outcomes after hospital discharge. *Journal of general internal medicine*, *19*(6), 624-631.

Table 1: Summary Statistics and Balance Tests for Non-Experimental and Experimental Schools

	Non Experimental Mean	Experimental vs. Non Exp. Sample
	(1)	(2)
Female	0.491	0.005
		(0.006)
White	0.102	-0.088***
		(0.018)
Black	0.179	0.186***
		(0.049)
Hispanic	0.661	-0.050
		(0.054)
Asian	0.045	-0.039***
		(0.008)
Other Race	0.013	-0.008***
		(0.002)
Economically Disadvantaged	0.788	0.147***
		(0.029)
Limited English Proficiency	0.444	-0.042
		(0.039)
Special Education	0.067	0.008
		(0.005)
Gifted and Talented	0.241	-0.078***
		(0.019)
Baseline Attendance Rate	97.401	-0.479***
		(0.151)
Baseline Behavioral Incidents	0.039	0.029***
		(0.008)
Pre-treatment Std. Math Score (STAAR & Stanford)	0.130	-0.443***
		(0.059)
Pre-treatment Std. Reading Score (STAAR & Stanford)	0.086	-0.312***
		(0.056)
Number of Clusters (schools)	124	174
Number of Students	32,866	45,581

Notes: The table describes summary statistics and balance tests for pre-treatment characteristics. Column (1) presents means for students attending grades 3 through 5 outside the experimental sample in 2013-2014. Column (2) presents the difference between attending grades 3 though 5 in the experimental sample versus the non experimental sample. This difference is estimated using an OLS regression of each pre-treatment characteristic on an indicator for being assigned to the experimental group. Standard errors, reported in parentheses, are clustered at the level of the school at time of treatment assignment. *. **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively. We use State of Texas Assessments of Academic Readiness (STAAR) test scores for students in grades 3-5 and Stanford 10 test scores for students in grades K-2. The last two rows provide summary statistics for an aggregated baseline score taking STAAR tests for students in grades 3-5 and Stanford 10 for students in grades K-2.

Table 2: Student Baseline Characteristics

	Control Mean	Treatment vs Control
	(1)	(2)
Female	0.492	0.005
		(0.007)
White	0.012	0.006
		(0.006)
Black	0.367	0.009
		(0.048)
Hispanic	0.608	-0.012
		(0.050)
Asian	0.008	-0.003
		(0.003)
Other Race	0.005	0.001
		(0.001)
Economically Disadvantaged	0.933	0.003
		(0.012)
Limited English Proficiency	0.391	0.013
		(0.033)
Special Education	0.078	-0.004
		(0.006)
Gifted and Talented	0.169	-0.010
		(0.013)
Baseline Attendance Rate	97.018	-0.227
		(0.159)
Baseline Behavioral Incidents	0.056	0.027***
		(0.009)
Pre-treatment Std. Math Score (STAAR & Stanford)	-0.304	-0.015
		(0.035)
Pre-treatment Std. Reading Score (STAAR & Stanford)	-0.223	-0.009
		(0.026)
Frac. Specialized Teachers (Survey Response)	0.542	0.357***
		(0.004)
Number of Clusters (schools)	25	50
Number of Students	6,019	12,715

Notes: The table describes summary statistics and balance tests for pre-treatment characteristics. Column (1) presents means for students attending grades 3 through 5 in a control school in 2013-2014. Column (2) presents the difference between attending grades 3 though 5 in a treatment school versus attending in a control school. This difference is estimated using an OLS regression of each pre-treatment characteristic on an indicator for being assigned to the treatment group controlling for matched pair fixed effects. Standard errors, reported in parentheses, are clustered at the level of the school at time of treatment assignment. *. **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively. We use State of Texas Assessments of Academic Readiness (STAAR) test scores for students in grades 3-5 and Stanford 10 test scores for students in grades K-2. We present summary statistics for an aggregated baseline score taking STAAR tests for students in grades 3-5 and Stanford 10 for students in grades K-2. For the last row we use a survey response to calculate the fraction of specialized teachers in treatment versus control schools. See Data Appendix for a detailed construction of all variables.

Table 3: The Effect of Treatment on State Test Scores

	ITT			2SLS (Ever)			2SLS (Years)		
	2014	2015	Pooled	2014	2015	Pooled	2014	2015	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Math	-0.062**	-0.033	-0.051*	-0.070**	-0.039	-0.058*	-0.073**	-0.021	-0.042*
	(0.031)	(0.035)	(0.028)	(0.034)	(0.040)	(0.032)	(0.036)	(0.021)	(0.023)
Reading	-0.057**	-0.018	-0.041	-0.064**	-0.021	-0.046	-0.067**	-0.011	-0.034
	(0.026)	(0.030)	(0.027)	(0.028)	(0.034)	(0.030)	(0.030)	(0.018)	(0.021)
Observations	11,266	10,683	21,949	11,266	10,683	21,949	11,266	10,683	21,949
Average Years				0.891***	0.865***	0.878***	0.850***	1.608***	1.218***
Of Treatment				(0.009)	(0.011)	(0.009)	(0.010)	(0.023)	(0.017)

Notes: This table presents estimates of being enrolled in or attending a treatment school on STAAR math and reading test scores. Here treatment is defined as attending a treatment school as the last school in 2012-2013. The sample is restricted each year to those students who are attending grades 3 through 5 and have both valid math and reading scores. Columns (1), (2), and (3) report Intent-to-Treat (ITT) estimates. Columns (4), (5), and (6) report 2SLS estimates and use treatment assignment as an instrument for having ever attended a treatment school. Columns (7), (8), and (9) report 2SLS estimates and use treatment assignment to instrument for the number of years spent in a treatment school. Columns (1), (4) and (7) use 2013-2014 score as the outcome variable. Columns (2), (5) and (8) use 2014-2015 score as the outcome variable. Columns (3), (6) and (9) use scores from both 2013-2014 and 2014-2015 as the outcome variable. The dependent variable in all specifications is state test score, standardized to have a mean of zero and standard deviation one by grade and year. All specifications adjust for the student-level demographic variables summarized in Table 2, student-level math and reading scores (3 years prior to 2013-2014) and their squares, and indicator variables for taking a Spanish baseline test. All specifications have grade year, and matched-pair fixed effects. Average years of treatment provides the first stage coefficient of instrumenting treatment with 2SLS (Ever) or 2SLS (Years) treatment variable. This number can be used to scale the ITT estimate into other estimates i.e. dividing the 2014 ITT estimate with 0.891 produces the 2014 2SLS (Ever) estimate. Standard errors (reported in parentheses) are clustered at the level of the school at time of treatment assignment. *, *, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table 4: Sensitivity Analysis or Extension of the Basic Model

	Math	*p-value*	Observations	Reading	*p-value*	Observations
	(1)	(2)	(3)	(4)	(5)	(6)
Full Sample	-0.051*	0.079	21,949	-0.041	0.131	21,949
	(0.028)			(0.027)		
Panel A: Demographics						
Male	-0.056**	0.046	10,779	-0.041	0.124	10,779
	(0.027)			(0.026)		
Female	-0.041	0.214	11,106	-0.039	0.197	11,106
	(0.033)			(0.030)		
Black	-0.031	0.212	7,534	-0.045**	0.038	7,534
	(0.025)			(0.021)		
Hispanic	-0.055	0.168	13,918	-0.033	0.381	13,918
	(0.039)			(0.038)		
Economically Disadvantaged: Yes	-0.044	0.140	20,522	-0.033	0.232	20,522
	(0.030)			(0.028)		
Economically Disadvantaged: No	-0.101***	0.007	1,363	-0.080***	0.009	1,363
	(0.036)			(0.029)		
LEP: Yes	-0.051	0.300	9,460	-0.044	0.409	9,460
	(0.049)			(0.052)		
LEP: No	-0.056**	0.020	12,425	-0.037**	0.041	12,425
	(0.023)			(0.018)		
Special Education: Yes	-0.156***	0.008	667	-0.199***	0.000	667
	(0.056)			(0.046)		
Special Education: No	-0.047	0.108	21,218	-0.038	0.175	21,218
	(0.029)			(0.027)		

Table 4: Sensitivity Analysis or Extension of the Basic Model

	Math	*p-value*	Observations	Reading	*p-value*	Observations
	(1)	(2)	(3)	(4)	(5)	(6)
Full Sample	-0.051*	0.079	21,949	-0.041	0.131	21,949
	(0.028)			(0.027)		
Panel B: Prior Achievement						
Gifted: Yes	-0.024	0.528	3,905	-0.019	0.649	3,905
	(0.038)			(0.041)		
Gifted: No	-0.066**	0.031	17,980	-0.054**	0.047	17,980
	(0.030)			(0.027)		
Baseline Test Tercile: T1	-0.012	0.647	8,696	-0.003	0.916	8,214
	(0.026)			(0.024)		
Baseline Test Tercile: T2	0.013	0.699	7,291	0.013	0.668	7,751
	(0.032)			(0.029)		
Baseline Test Tercile: T3	-0.023	0.404	5,525	0.024	0.430	5,629
	(0.027)			(0.030)		
Panel C: Teacher Characteristics						
Teacher Age Tercile: T1	-0.174***	0.000	1,926	-0.208***	0.000	2,070
	(0.041)			(0.052)		
Teacher Age Tercile: T2	0.107**	0.033	1,741	0.072	0.413	1,744
	(0.049)			(0.087)		
Teacher Age Tercile: T3	0.025	0.695	1,696	-0.077	0.330	1,649
	(0.063)			(0.078)		
Difference in TVA Tercile: T1	-0.185***	0.000	3,896	-0.104**	0.031	4,562
	(0.031)			(0.047)		
Difference in TVA Tercile: T2	-0.021	0.718	3,537	-0.107***	0.002	2,793
	(0.059)			(0.033)		
Difference in TVA Tercile: T3	0.008	0.860	3,833	0.035	0.314	3,911
	(0.044)			(0.034)		

Notes: This table presents pooled ITT estimates of the effect of being enrolled in a treatment school on STAAR math and reading test scores in different subgroups of the sample. Panels A and B split the sample according to student characteristics while Panel C splits the sample according to teacher characteristics. For teacher age, the entire sample of teachers from HISD district were divided into terciles based on their ages. For teachers' difference in Teacher Value Added (TVA), we calculated the difference between the TVA of the subject that a treatment teacher actually taught in 2013-2014 and the average TVA across all subjects that the treatment teacher used to teach before. These differences were averaged per treatment school. Then, treatment schools were divided into terciles based on the mean difference in TVA. Control schools received the same tercile as the treatment school in their own matched pair. All specifications follow the pooled specification from Table 3. All standard errors, located in parentheses, are clustered at the level of the school at time of treatment assignment. *. **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table 5: The Effect of Treatment on Outcomes, School-Level Regressions

	2014	2015	Pooled
	(1)	(2)	(3)
Panel A: Student Achievement			
Math	-0.096***	-0.085***	-0.083***
	(0.002)	(0.003)	(0.002)
Reading	-0.081***	-0.057***	-0.071***
	(0.002)	(0.003)	(0.002)
Observations	11,266	10,683	21,949
Panel B: Alternate Outcomes			
Attendance (in years)	-0.002***	-0.003***	-0.003***
	(0.000)	(0.000)	(0.000)
Observations	12,713	11,667	24,380
Behavioral Incidents	0.016***	–	0.016***
	(0.001)	(–)	(0.001)
Observations	12,713	–	12,713
Teacher Retention	-0.021	–	-0.021
	(0.042)	(–)	(0.042)
Observations	50	–	50

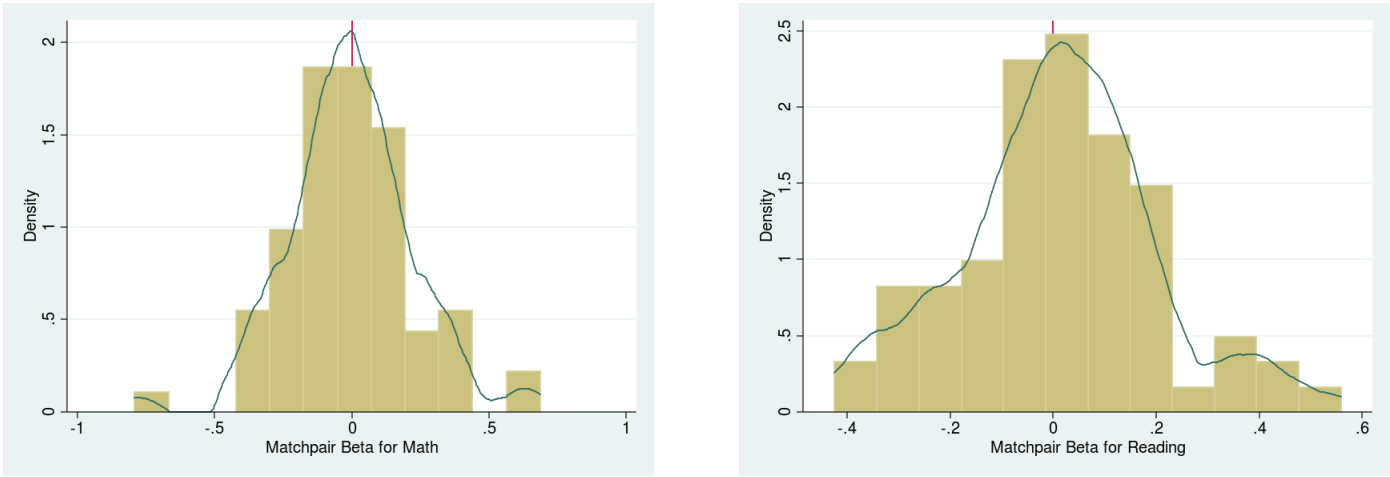
Notes: This table presents the estimates of being enrolled in a treatment school on school-level averages of STAAR test scores and other alternate outcomes. Panel A presents ITT estimates on math and reading scores. Panel B presents ITT estimates on average attendance rates (measured in years), average behavioral incidents and teacher retention. Teacher retention is calculated as the fraction of teachers retained between 2012-2013 and 2013-2014 per school. See Data Appendix for a detailed construction of all variables. Column (1) uses outcomes from 2013-2014, column (2) uses outcomes from 2014-2015, and column (3) uses outcomes from both years. The specifications follow the main OLS specification from Table 3 at the school-level rather than the individual level. The mean of demographic controls is taken at the school level. The school-level mean of student's 2010-2011, 2011-2012, and 2012-2013 test scores are taken for controls. School level means of attendance rates and behavioral incidents are included when the outcome variable is attendance rate and behavioral incidents, respectively. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table 6: The Effect of Treatment on Stanford 10 Test Scores

	ITT			2SLS (Ever)			2SLS (Years)		
	2014	2015	Pooled	2014	2015	Pooled	2014	2015	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Grades 3 - 5									
Math	-0.061**	-0.004	-0.040	-0.068**	-0.005	-0.046	-0.072**	-0.002	-0.034
	(0.027)	(0.032)	(0.027)	(0.030)	(0.037)	(0.030)	(0.032)	(0.020)	(0.022)
Reading	-0.061**	-0.036*	-0.056**	-0.069**	-0.042*	-0.064**	-0.072**	-0.023*	-0.048**
	(0.027)	(0.021)	(0.023)	(0.030)	(0.023)	(0.026)	(0.031)	(0.013)	(0.019)
Observations	11840	9424	21264	11840	9424	21264	11840	9424	21264
Panel B: Grades 1 - 2									
Math	0.013	-0.074*	-0.008	0.015	-0.088*	-0.010	0.015	-0.049*	-0.009
	(0.037)	(0.042)	(0.034)	(0.041)	(0.049)	(0.039)	(0.044)	(0.027)	(0.035)
Reading	-0.021	-0.071	-0.030	-0.024	-0.085	-0.035	-0.025	-0.047	-0.031
	(0.034)	(0.054)	(0.032)	(0.038)	(0.062)	(0.036)	(0.040)	(0.035)	(0.032)
Observations	8192	2317	10509	8192	2317	10509	8192	2317	10509
Panel C: All Grades									
Math	-0.038	-0.006	-0.030	-0.043	-0.008	-0.035	-0.045	-0.004	-0.027
	(0.027)	(0.030)	(0.026)	(0.030)	(0.034)	(0.029)	(0.031)	(0.019)	(0.023)
Reading	-0.052**	-0.035	-0.049**	-0.059**	-0.041	-0.056**	-0.062**	-0.022	-0.044**
	(0.024)	(0.022)	(0.023)	(0.027)	(0.025)	(0.026)	(0.028)	(0.014)	(0.021)
Observations	20032	11741	31773	20032	11741	31773	20032	11741	31773

Notes: This table presents estimates of the effect of being enrolled in a treatment school on Stanford 10 test scores. The sample and specification is identical to that used in Table 3. All standard errors, located in parentheses, are clustered at the level of the school at time of treatment assignment. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Figure 1: Matchpair Specific Treatment Effects



Notes: These figures plot a kernel density curve for matchpair by grade treatment coefficients. Treatment coefficients are obtained by regressing the math STAAR score or reading STAAR score on student demographics, baseline test scores and year fixed effects for each matched pair and grade. The figures also display an xline at 0 for comparison.