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

Roland G. Fryer, Jr.
Harvard University and NBER

August 2017

Abstract

Starting in the 2013-2014 school year, I conducted a randomized field experiment in forty-six 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 encouraging schools to specialize their teachers on student achievement is -0.11 standard deviations per year on a combined index of math and reading test scores. Students enrolled in special education and those with less experienced 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, 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, Chad Syverson, the Editor and three anonymous referees for helpful comments and suggestions. Meghan Howard Noveck and Rucha Vankudre provided exceptional project management oversight. Tanaya Devi and Hannah Ruebeck provided particularly exceptional research assistance. 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 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 may reduce transition times from one task to the next during which productivity is lost. Lastly, division of labor allows individuals to focus their full

[1] Although Adam Smith popularized the notion of division of labor through his theory of 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.

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, five consistently use specialized teachers in classrooms in elementary schools. Of those five, all begin specializing teachers during or after grade 3. 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.

While economists have speculated about the potential benefits and low cost of a policy that assigns teachers to teach subjects in which they are most effective (e.g. Jacob and Rockoff 2011), existing evidence for the policy is correlational. For example, the estimates in Condie, Lefgren, and Sims (2014) suggest that there is useful variation in teachers' effectiveness across subjects and that using measures of value-added to assign teachers to teach the subject in which they are more effective could increase student achievement by up to 0.08σ per year.

Additionally, interest in elementary school teacher specialization is growing in the US. The Common Core State Standards, introduced in 2010 and currently utilized in 40 states, require students to achieve at higher levels and to demonstrate new skills, in many cases requiring more in-depth content knowledge from those teaching even the youngest children (Gewertz 2014). Even before the advent of the Common Core, teacher specialization in elementary schools was considered by many school leaders as a potential way to better prepare teachers to meet accountability standards in the era of high-stakes testing.

Starting in the 2013-2014 school year, I conducted a randomized field experiment in forty-six traditional public elementary schools in Houston, Texas, designed to test the potential productivity benefits of teacher specialization and shed light on what mechanisms

may 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 90% of grade-language cells. Based on this availability, teams of teachers were designated within schools, grades, and languages taught. After reviewing schools' departmentalization plans, I recommended further changes in teaching assignment for less than five percent of the cases and half of these recommendations were accepted; the teacher assignments in which principals did not accept these recommendations were all judgment calls for which I deferred. Control teachers continued the status quo. If there were any specialized teachers in control schools, they were kept as such.

As expected, assignment to treatment significantly increased the degree of specialization within a school. Using both administrative and survey data, I show that teachers in treatment schools were approximately sixty percent more likely to be specialized (i.e., to not teach both math and reading, to teach three or fewer subjects, or to self-report teaching several classes of different students in 1-3 subjects) in year one and approximately fifty percent more likely to be specialized in year two, relative to teachers in control schools. Consistently, teachers taught, on average, approximately 30 percent (16 percent) more students than control teachers in year one (year two). Among teachers who were not already

[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. Details are provided in Section III.

specialized in the year prior to treatment, assignment to treatment more than doubled the probability that a teacher was specialized in the first year of the experiment.

Additionally, I construct a measure of specialization for each teacher that takes into account the quality of fit of teachers to the subjects they teach as well as the number of subjects they teach, using both the "ideal" and actual assignment of teachers to subjects. Dividing the actual degree of specialization by the "ideal" degree of specialization yields the percent of potential specialization that is being utilized by each grade-language cell. Treatment increases this measure by 26 percentage points in year one (control mean = 0.43) and 13 percentage points in year two (control mean = 0.53). Pooled over both years of treatment, the increase in the percent of potential specialization that is utilized in treatment schools is largest for the youngest grades and decreases substantially as children age.

The effects of the experiment on student achievement are surprisingly *inconsistent* with the positive effects of specialization typically known to economists. In the first year of the experiment, the (Intent-to-Treat) impact of encouraging elementary schools to specialize their teaching staff was -0.12σ (0.05) on an index of high-stakes test scores and -0.11σ (0.05) on an index of low-stakes test scores. Both indices are summed effects on math and reading; to get an average effect per subject divide the estimates by two. In the second year of treatment, treatment effects were -0.09σ (0.05) on both high- and low-stakes test scores. Pooled across years, students in treatment elementary schools score 0.11σ (0.04) *lower* on high-stakes exams and 0.10σ (0.04) *lower* on low-stakes exams, per year, relative to students in control elementary schools.

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 –0.30σ (0.09) on high-stakes tests and -0.22σ (0.07) on low-stakes tests. For non-special education students, the impact of treatment is -0.10σ (0.04) on high- and low- stakes tests. The p-value on the difference is 0.04 on high-stakes tests and 0.09 on low-stakes tests. Students who were taught by less experienced teachers also demonstrated large negative treatment effects. Students with more experienced teachers had either less negative or zero effect of treatment.

Beyond test scores, students in treatment schools, on average, have 1.12 times as many serious behavioral infractions and attend 0.36 fewer days of school per year than students attending control schools. The attendance effect is statistically significant.

I argue that familiarity with student type explains at least 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. On average, treatment teachers taught 26% *fewer* subjects than control teachers. A cost is that treatment teachers had, on average, 23% *more* total student contacts than teachers in control schools – raising the costs of individually tailoring pedagogy.[3] 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. Moreover, treatment teachers were much less likely to report an increase in job satisfaction or performance than teachers in control schools. 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 some potential 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 discusses robustness checks of the main results. Section VII 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 define key variables used in the empirical analysis.

[3] For teachers on the margin (those who were not specialized in the year previous to treatment), treatment decreased the number of subjects taught by 30% and increased the number of student contacts by 30% over both years of treatment.

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. Intuitively, these channels may operate differently for students in different grade levels.

In school systems across developed countries, teacher specialization consistently increases as students age. The average OECD country specializes teachers starting in grade 6. This may be, for example, because in younger grades the increased cost of tailoring pedagogy outweighs the benefits from teachers' increased subject-specific knowledge whereas the opposite is true in older grades where subject material is more complex.

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 teachers are responsible for, allowing them to focus more energy on lesson planning and other subject-specific investments.[4] Third, some argue that specialization increases teacher retention due to reduced workload and reduced likelihood of teaching an unfamiliar subject.[5] 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 lack of economies of scale, coordination costs, and principal agent problems between workers.

[4] Teachers could also use the additional time for increased leisure.

[5] Teacher retention was not significantly different between treatment and control schools. Appendix Table 8 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.

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 teachers have 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.[6] Increased transition times between classes can also decrease 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.

***

I cannot credibly identify the separate impact of each of these potential costs and benefits. Moreover, it is also possible that the gains from better teaching due to teacher specialization are diluted by reduced effort from students, parents, or school principals in response to better classroom instruction if teaching quality and student, parent, or principal effort are strategic substitutes. Of course, the opposite may also come true – the other agents involved in educational production could increase their effort in response to better teaching, augmenting the gains from better teaching (i.e. strategic complements).

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. Although I present results that attempt to unearth the mechanisms underlying the net impact of teacher specialization, this discussion is (necessarily) more speculative.

---

[6] 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.

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 through which 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 expected total student achievement.

Payoffs

I 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: $\sum_{j=1}^{N} Y_j = f(e_j, \theta_j, H) - (\alpha_j - x)^2$.[7]

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:

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

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 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'

[7] One can allow 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.

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 $e^* \rightarrow \frac{\partial f}{\partial e_j} - \frac{\delta k}{\delta e_j} = 0$.

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, $T \to 0$.[8]

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 increases 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.[9]

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

[8] 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.

[9] 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.

To begin the field experiment, I followed standard protocol. First, I garnered support from the district superintendent and other key district personnel. The district then provided a list of 62 K-5 elementary schools that were eligible for the teacher specialization experiment.[10] I removed sixteen of these schools because either they were part of another experiment (e.g. Fryer 2014) or because their particular school model was antithetical to the notion of teacher specialization (e.g. Montessori). The final experimental sample consists of forty-six schools – twenty-three treatment and twenty-three 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. Treatment schools then sent in specialization plans along with a written justification for each plan.[11] 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).[12] Control schools were not told they were ever considered for participation in the experiment.

If schools were hesitant about implementing part of the specialization plan, they were required to submit written justification to HISD's Chief Academic Officer. This might include cases in which teachers did not want to participate and lobbied school principals to reconsider the specialization plans. There was no incentive offered to treatment schools to follow the specialization plans. How schools decided to manage the organizational changes required due to specialization was left to each principal, with my input. Anecdotally, in some schools, teachers rotated classrooms and in others, students would move between rooms

[10] Schools were not alerted that they were being considered for participation in an experiment. 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. HISD focused on low-performing schools out of a desire to improve student outcomes, not because they thought that specialization would be more effective in these schools. All eligible schools offered only grades K-5; combined schools that offered grades K-6 or K-8 were excluded from participation.

[11] Treatment schools were told that they would be altering their schedules and specializing their teachers on August 8, 2013. Schools submitted their specialization plans for approval on August 22, 2013.

[12] Jacob and Lefgren (2008) show that principal observations are generally accurate in identifying teachers who are the most or least effective at teaching math or reading but are less accurate in judging teachers in the middle of the distribution. Rockoff et al. (2010) show that both principals' evaluations of teachers and objective measures of teachers' performance are sensitive to teachers' subject-specific skills – i.e. principals can distinguish between strong math teachers and strong reading teachers.

throughout the day. Unfortunately, I do not have information on changes to these organizational structures within schools or estimates of potential instructional time lost to transition time between classrooms or teachers.

Teachers were notified of the changes in teaching assignments as soon as they returned to school from summer vacation. While this may seem like late notice, no teacher was assigned to teach a subject they had not taught previously and it is exceedingly difficult to contact teachers over the summer.[13] Moreover, this also prevented teachers and/or students from switching schools in response to changes in teaching assignments in treatment schools.

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 90% of grade-language cells. Based on this availability, teams of teachers were designated within schools, grades, and languages. 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.[14]

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 (1-3) 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 (LEP) students to English-only instruction

[13] One might argue that changing teachers' schedules and the ways that teachers interact with students could be driving the negative results of specialization, and that these negative effects may be reduced or eliminated over time. In this case, two years may not be enough time to observe the effects of the treatment. Arguing against this, I find starkly negative effects for first-year teachers, suggesting that it is specialization itself and not the disruption that occurred because of the transition from self-contained to specialized classrooms that is negatively affecting student achievement.

[14] Control schools continued the status quo. In traditional "self-contained" classrooms (i.e. non-specialized classrooms), students are instructed by one teacher with the same group of classmates for all subjects.

in grades 4 and 5. The English as a Second Language (ESL) program is offered to those students with a home language other than Spanish – students in the ESL program are instructed in English using ESL methods appropriate for each subject. In grades 1-3, 25 percent of grade-language cells had only one teacher if the language of instruction was Spanish, whereas all of the grade-language cells had at least two teachers if the language of instruction was English. Therefore, for the youngest children, the potential for specialization was much lower for LEP students than for non-LEP students. In grades 4 and 5, teachers could be specialized across languages due to the structure of the transitional bilingual classrooms, where the language of instruction is determined by content area – in fourth grade, language arts, reading and math are taught in Spanish and science, social studies, and English literature are taught in English and in fifth grade language arts, reading, math, and science are taught in English and social studies and Spanish literature are taught in Spanish. Due to this complexity, I consider these differences by language group in Section V.

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 my 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, I 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.

Table 1, Columns (1)-(3) describe differences between experimental elementary schools (both treatment and control) and all other elementary schools in HISD across a set of covariates gleaned from administrative data. The descriptive differences between experimental and non-experimental 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, less likely to be gifted, and have lower pre-treatment test scores in math and reading.

Teachers have lower TVA in both math and reading, on average.[15] See Online Appendix B for details on how each variable was constructed.

Thus, the results estimated are likely more applicable to urban schools with high concentrations of minority students. Overall, the participating versus non-participating sample is very unbalanced at the school, teacher, and student levels (p-values on the joint F-tests are all 0.000). However, the pre-treatment degree of specialization (actual and ideal) is similar between participating and non-participating schools at both the school and teacher level, suggesting that the leadership of HISD was not selecting schools that were better or worse prepared to departmentalize their teaching staff.

IV. Data, Research Design, and Econometric Framework

Data

I 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 or Iowa Test of Basic Skills (ITBS) subject scores in math, reading, science, and social studies for students in first through fifth grades. Behavior data records student behavioral incidents resulting in a serious disciplinary action such as a suspension or an expulsion.

Additional data files link students to their teachers in each subject. I have HISD data spanning the 2010-2011 to 2014-2015 school years. To supplement HISD's administrative data, I also collected data from a survey administered to teachers at the end of the 2013-2014 school year, described below.

[15] I use the official TVA calculated by the district for any teacher for whom it is available. Due to a limited sample of teachers with HISD-calculated TVA measures, I fill this measure with an author-calculated measure of teacher effectiveness in 2012-13 for teachers who are missing the HISD-calculated measure. This increases the number of teachers in experimental schools with measures of teacher effectiveness by 130 percent in math and 140 percent in reading. Both the HISD- and author-calculated measures are separately standardized over the entire district to have a mean of zero and standard deviation one. The correlation between the two measures is 0.8 in math and 0.6 in reading. The author-calculated measure controls for student demographics and previous year test scores – for details on its construction, see Appendix B. Results are similar using only HISD-calculated TVA (available upon request).

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.[16] 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. I use a student's first score unless it is missing.[17]

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 this 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.[18]

Houston is one of a handful of cities in the US that voluntarily 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. In the 2013-2014 school year and years previously, HISD administered the Stanford 10. In the 2014-2015 school year, HISD administered the Iowa Test of Basic Skills (ITBS). Both tests are aligned with standards set by the National Council of Teachers of Mathematics, the National Council of Teachers of English, the International Reading Association, and the National Assessment of Educational Progress. Both assessments test math and reading in grades K-5; the Stanford 10 also tests science and social science in grades 3-5 and the ITBS tests science and social studies in grades 1-5.

---

[16] Sample tests can be found at http://www.tea.state.tx.us/student.assessment/released-tests/.

[17] Using their retake scores, when the retake is higher than their first score, does not significantly alter the results. See Appendix Table 13.

[18] 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, I did not use them for the main analysis. I did, however, investigate whether treatment influenced whether or not a student takes a standard or non-standard test (see Appendix Table 8) and the effect on STAAR-M test scores (see Section 5). The 2014 spring STAAR administration was the last to offer the Modified exam. In 2015, students with special needs took the new Accommodated version of the test (STAAR--A), which is comparable to the regular version of the test but administered online with special accommodations. Students taking STAAR--A must meet the regular STAAR performance standards.

I 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 I include in all regressions (unless otherwise noted), and are also referred to throughout the text as "pre-treatment test scores." I also include one indicator variable for each pre-treatment test score that takes on the value of one if that test score is a Spanish version test and zero otherwise, and an indicator that is a one if a student is missing each pre-treatment test score. Students are not required to have pre-treatment scores to enter the sample. Pre-treatment scores are high-stakes STAAR test scores in math and reading – the scores that random assignment was designed to balance (details below).

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.[19]

I use data linking students to their teachers in each subject to develop teacher-level measures of specialization in 2012-2013 and 2013-2014. The data allow for multiple measures of specialization, including whether a teacher teaches both math and reading, the number of subjects taught, the number of grades taught, and the total number of students taught. The same data are used to determine whether a teacher teaches in English or Spanish. I include all teachers working in a school except those who teach Special Education classes or are Special Education specialists, who would not have been included in the experiment and who often work with students in a different capacity than teachers of traditional classes.

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 first year of

[19] 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.

treatment). The data from the survey includes questions about lesson planning, relationship with students and interaction with parents and guardians of students. In addition, measures of specialization were constructed from survey responses that match those culled from the administrative data. 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. 395 (78% response rate) treatment teachers and 315 (72% response rate) control teachers completed the survey. See Online Appendix B for details on the construction of outcomes used from the survey and the administrative dataset.

Research Design

To partition the set of schools provided by the district into treatment and control, I used a matched-pair randomization procedure. Recall, forty-six schools entered the experimental sample from which I constructed twenty-three 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 high-stakes state test scores. First, the set of forty-six schools were ranked by the sum of their mean reading and math test scores in the previous two years. Then, I designated every two schools from 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 (4)-(6) of Table 1 display descriptive statistics on school, teacher, and individual student characteristics of all HISD students enrolled in a final-sample experimental school in first through fifth grade with valid outcome high- or low-stakes test scores in both math and reading after the first year of treatment. Columns (4) and (5) provide the mean for each variable for the control and treatment group, respectively. Column (6) provides the p-value on the difference between the treatment and the control group, which I estimate by regressing the variable on a treatment indicator. See Online Appendix B for details on how each variable was constructed.

Panels A and B include measures of teacher specialization in treatment and control schools in the year prior to treatment. There are no significant differences in any of the measures of pre-treatment specialization. Overall, the treatment and control groups are balanced on every observable characteristic at the school level – the unit of randomization –

making inference relatively straightforward (although the p-value on the joint F-test is 0.063). In Panels B and C, no individual observable variable is significantly different between treatment and control; the p-value on the joint F-test at the student level is 0.721.

Econometrics

To estimate the causal impact of teacher specialization on outcomes, I estimate both intent-to-treat (ITT) effects and Local Average Treatment Effects (LATEs). For individual i 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, and 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-three treatment and twenty-three control schools in the 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 encouraging schools to specialize their teaching staffs, 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. control) if they were in a treatment (resp. control) school in the pre-treatment year and not in an exit grade in the pre-treatment year (e.g. 5th grade). Students in both Spanish- and English-instruction programs were potentially treated and are included in the analysis. All student mobility after treatment assignment is ignored. Note that Equation (4) is estimated on first through fifth graders in each year of treatment and treatment assignment was determined in the pre-treatment year. Thus, 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 these estimates will be biased upwards (resp. downwards) – even under random assignment. I find that 8.75% of treatment student observations are missing a state test score in either year relative to 9.54% of control students, a difference of 0.79%. Thus, despite attrition rates being around 9.12%, 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 Section VI.

I also estimate three different Local Average Treatment Effect (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. The second LATE parameter is estimated through a two-stage least squares regression of student achievement on the intensity of treatment. More precisely, I define *YEARS* as the fraction of years a student is present at a treatment school. The final LATE parameter, *DEGREE*, is the percent of potential specialization being utilized in a student's grade-language cell.

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

Proof of Treatment

I begin the analysis of the experiment by describing how treatment affected eight measures designed to capture the degree of teacher specialization in treatment schools relative to control schools. I estimate a linear regression for each measure of teacher specialization on an indicator for working in a treatment school and matched-pair fixed effects. Standard errors are clustered at the school level. Each measure addresses various aspects of specialization. Measures include: indicators for whether a teacher teaches both math and reading, indicators for whether a teacher teaches fewer than 2 or 3 subjects and a linear measure of how many subjects a teacher teaches, whether a teacher considers themselves "departmentalized" (i.e. specialized), and the total number of students a teacher comes into contact with daily.

A final summary measure calculates the percent of potential specialization utilized by a school, which accounts for the fact that schools may be differentially able to specialize, given certain constraints (e.g. knowledge of the comparative advantage of their existing

teaching staff, the number of teachers in each grade-language cell, etc.). It additionally considers the "quality of fit" of each teacher to teach the subjects that he or she is assigned to teach, using measures of TVA. For details on the construction of this measure, see Online Appendix B.

Table 2A demonstrates – using both administrative and survey data – that schools randomly chosen for treatment increased their level of specialization. This effect is remarkably consistent across each measure of teacher specialization.

At the control mean, 48 (55) percent of elementary school teachers do not teach both math and reading in year one (two); the identical number in treatment schools is 85 (82) percent. The relatively high rates of teacher specialization in control schools (approximately 50 percent of teachers) may seem notable.[20] However, in a nationally representative survey of public school teachers in 2012, 11 percent of public school elementary teachers were purely specialized, 32 percent were partially specialized, and 57 percent were purely non-specialized (NCES Schools and Staffing Survey).

One potential explanation for the higher-than-average baseline rates of teacher specialization in HISD is the decentralized – school-based budgeting – structure of the district. In particular, schools can make changes to teacher assignments and school schedules without the approval of district administrators. This explanation may be particularly plausible given the recent increase in national interest in elementary school specialization in response to high-stakes accountability initiatives and the Common Core State Standards, both of which require more in-depth subject knowledge from elementary school teachers (Gewertz 2014). Rates of teacher specialization (in non-experimental or control schools) are not correlated with any observable school characteristic (average teacher TVA or experience, student demographics, etc.).

Treatment reduced the average number of subjects taught from 2.9 to 2.0 in year one and from 2.7 to 2.1 in year two. In addition, the average number of students that a teacher taught increased from 38 students in control schools to 49 in treatment schools in year one and from 42 to 49 in year two. Finally, treatment increases the percent of potential specialization utilized by the average school by 26 percentage points (control mean = 0.43) in year one and 13 percentage points (control mean = 0.53) in year two.

[20] Restricting the sample to matched pairs with control schools that have teacher specialization rates under fifty percent yields results similar in magnitude and significance to these main results (See Appendix Table 14).

Table 2B provides similar estimates for the subsample of marginal teachers whose assignments were likely to be affected by their school's assignment to treatment. A teacher is considered marginal if she taught both math and reading in the year previous to treatment or is new to the district. As expected, marginal teachers were much more likely to be affected by treatment, relative to the marginal teachers in control schools.

Note: the increase in the degree of specialization in year one is remarkably consistent both across different measures of specialization and across the administrative data and survey responses, providing some assurances that the experiment produced the desired behavioral response from schools. The estimates using survey-based measures of specialization to determine proof of treatment are robust to standard bounding procedures (Lee 2009) despite the ten percentage point differential in treatment and control response rates to the survey (see Appendix Table 1).

Finally, treatment schools were more consistently specialized whereas control schools have more varied teacher specialization rates (Figure 1). In year one, all treatment schools had at least 60 percent of teachers specialized (defined as not teaching both math and reading). In year two, half of the treatment schools had at least 80 percent of teachers specialized.

The Impact of Treatment on State Test Scores

Table 3 presents a series of ITT estimates of the impact of encouraging schools to specialize their teachers on summed math and reading high- and low-stakes test scores. Recall, high-stakes test scores are math and reading state test scores that are used for accountability purposes in the state of Texas. Low-stakes test scores are math and reading test scores from the Stanford 10 (in 2013-14) and the Iowa Test of Basic Skills (in 2014-15). To get an average effect per subject, divide the estimates by two. Test scores are normalized to have mean 0 and standard deviation 1 across the entire school district by subject, grade and year, so treatment effects are presented in standard deviation units. Standard errors, clustered by school and year, are in parentheses below each estimate along with the number of observations and the control mean. All regressions include grade, year and matched-pair

fixed effects and the controls described above.[21] The high-stakes sample includes students in grades 3-5 and the low-stakes sample includes students in grades 1-5.

Columns (1) and (2) present ITT estimates of the impact of treatment on high-stakes test scores for the 2013-2014 and 2014-2015 school years, respectively. Columns (4) and (5) present analogous results for low-stakes test scores. In the first year of the experiment, the impact of encouraging schools to specialize their teachers was -0.12σ (0.5) on high-stakes tests and -0.11σ (0.05) on low-stakes tests. In the second year of treatment, treatment effects were a bit less negative (-0.09σ on both high- and low-stakes test scores) and no longer significant. Pooling across years, students in treatment elementary schools score 0.11σ (0.04) *lower* on high-stakes tests and 0.10σ (0.04) *lower* on low-stakes tests per year relative to students in control elementary schools. These results are robust to a range of alternative specifications and other tests, presented in the final rows of Table 3 and discussed in detail in the next section.

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. The remarkable consistency in the effects on high- and low-stakes test scores described above suggests that teacher specialization leads to actual losses in academic knowledge rather than simply decreasing teachers' efficiency at high-stakes test preparation.

Appendix Table 2 provides estimates for each high-stakes subject separately; Appendix Table 3 presents similar estimates for each low-stakes subject. High- and low-stakes math and reading are all negatively affected by teacher specialization in both years of treatment. In year one, effects on reading are slightly larger than effects on math. In year two, the high-stakes math effect is larger than the reading effect but the low-stakes reading effect is larger than the math effect.

Table 4 presents LATE estimates that scale the ITT results by student attendance in a treatment school. Columns (4)-(6) present LATE estimates for the cumulative effect of actually attending a treatment school. The average cumulative effect of attending a treatment school for at least one day in any of the school years is -0.12σ (0.04) on both high- and low-

[21]The treatment effects estimated by Equation (4) are consistent with those obtained from an identical model without demographic or baseline test controls; see Appendix Table 15.

stakes tests, per year, pooled over both years. 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 pooled estimates in Column (9) by two. The impact of teacher specialization is -0.09σ (0.03) on both high- and low-stakes tests, per year. Thus, at the end of the two-year experiment encompassing 18 school months, students attending treatment schools were approximately one to one and a half months behind students attending control schools – implying that encouraging elementary schools to specialize their teachers reduces production efficiency by almost 10 percent.

Overall, these results are surprisingly *inconsistent* with the positive effects of division of labor typically known to economists though, as Proposition 1 illustrates, they 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 investigate the data is to graph the distribution of treatment effects for each matched pair-grade cell, which is depicted in Figure 2. I control for demographic observables and baseline test scores by estimating equation (4) for each matched pair. I then collect the treatment coefficients from this equation and plot a kernel density curve for them. The results echo those found in Table 3. Out of 23 matched pairs, 15 (17) have negative results on high-stakes (low-stakes) test scores.

The variation in matched pair treatment effects is partially explained by differential levels of implementation of the teacher specialization policy. Figure 3 provides a visual representation of this classic "dose-response" relationship. For each matched pair, I estimate equation (4) with the percent of potential specialization utilized as the dependent variable – i.e. the matched pair first stage effect – and plot the regression-adjusted matched pair treatment effect against the regression-adjusted matched pair first-stage effect.

When estimated on the full sample of students in grades 1-5, the slope of the dose-response is strikingly negative in both years of treatment. Increasing the percent of specialization utilized from 0 to 100 percent *reduces* student achievement by 0.88σ in the first year and 0.82σ in the second year on low stakes tests (both p-values are 0.02). The relationship is negative but insignificant when estimated on the 3rd-5th grade sample; a significant fraction of upper elementary school classrooms were already specialized. Estimating this slope only for grade-language cells that were utilizing less than 50 percent of

potential specialization in the year before treatment yields an even more negative relationship. Conversely, the slope is flat when estimated for grade-language cells that were already utilizing more than 50 percent of potential specialization in the year before treatment. No other observable treatment school characteristics significantly predict matched pair treatment effects (Appendix Figure 1, Panel A).

Given the negative relationship illustrated in Figure 3, one can scale the ITT results by the exogenous increase in the degree of specialization induced by treatment to estimate the impact of increasing teacher specialization on student achievement. Appendix Table 4 presents these LATE estimates. Using the LATE specification, the impact of increasing the percent of specialization utilized from 0 to 100 percent is to *lower* test scores by 0.98σ (0.37) on the high-stakes exam in grades 3-5 year one and 1.03σ (0.72) in year two. In the low-stakes sample (grades 1-5), increasing the percent of specialization from 0 to 100 percent *lowers* test scores by 0.44σ (0.19) in year one and 0.81σ (0.42) in year two.[22]

The Impact of Treatment on Attendance and Behavior

To understand the impact of treatment on attendance or behavior, I estimate a regression specification identical to that used to estimate the effects on student test scores, with an added control variable that measures the outcome of interest (attendance or behavior) in the year prior to treatment. Consistent with the negative impact on test scores, treatment increases the number of student suspensions and decreases attendance rates, on average. Specifically, students in treatment schools are 1.13 times as likely to be suspended due to poor behavior per year and attend 0.36 lesser number of days of school per year. The results for attendance are statistically significant (Appendix Table 5).

Heterogeneous Treatment Effects

Table 5 presents ITT results exploring the sensitivity of the estimated treatment effects across various subsamples of the data. For the parallel LATE estimates, see Appendix Tables 6 and 7. Appendix Figure 1 plots analogous treatment effects over the distribution of continuous variables used to subsample the data, rather than using the above-below median cutoff. The negative effects of teacher specialization are remarkably robust across various

[22] These differences are driven by differences in samples (grades 3-5 versus grades 1-5), not by differential impacts on high- versus low-stakes test scores. Differences by grade level are further discussed below.

subsamples, though there is some evidence that certain students who are more likely to need individual attention (in a dial-setting sense) – e.g. students with special needs in regular classroom environments – do particularly poorly when teachers are specialized.

Recall, the key tradeoff of teacher specialization is between the benefit of having a teacher with increased subject-specific content mastery or lesson preparation and the cost of having a teacher with less information about students' idiosyncratic type – leading to worse targeting of students' needs.[23] If certain students benefit more from subject-specific content mastery (students in the gifted and talented program, say), or certain students have more specific learning needs (students in the special education program, for example), they may experience differential effects of specialization. Additionally, it is possible that there are groups for whom a teacher needs more signals to accurately determine their learning needs – for example, students who are of a different race than their teacher (Delpit 2006).

Consistent with the tradeoff described above, the coefficient on treatment for students with special needs is -0.30σ (0.09) and the effect for students without special needs is -0.10σ (0.04) on high-stakes exams; the respective estimates on low-stakes exams are -0.22σ (0.07) and -0.10σ (0.04).[24] The p-value on the difference is 0.037 for high-stakes test scores and 0.088 for low-stakes test scores. The coefficient on students who participate in the gifted and talented program is -0.04σ (0.06) and the effect for students who do not participate in the program is -0.13σ (0.04) on high-stakes exams; the respective estimates on low-stakes exams are -0.07σ (0.05) and -0.11σ (0.04). The difference is statistically significant for high-stakes test scores. Taken together, these results are consistent with the expectation that students who may have more specific pedagogical needs lose the most from teacher

[23] While there are other potential costs and benefits to teacher specialization (described in detail in Section II), other potential mechanisms are less consistent with the patterns evident in the data that are described in this section. That said, this discussion is merely suggestive.

[24] This is the effect for students with special needs who take the Standard STAAR test. In 2013-2014, half of the special needs students in the sample took the Modified version of the test (STAAR-M). These are students who require extensive classroom modifications and won't achieve grade-level proficiency. There was no treatment effect on exiting the main sample due to taking a STAAR-M exam (Appendix Table 8). The treatment effect on students with STAAR-M test scores (following the same specification as Table 3 but with standardized STAAR-M test scores as baseline test score controls) is -0.09σ (0.14) on high stakes scores and -0.10 (0.9) on low-stakes scores. Neither of these effects is significantly different from zero. These students could be more protected from the negative effects of inefficient dial setting if they are more likely to have aides or if teachers are more aware of their needs than they are of the needs of the students who have special needs but still take the regular state test.

specialization while those who may benefit from teachers' additional content mastery lose the least from teacher specialization.

Another potential dial-setting vulnerability is teacher experience. Teachers with more experience may more easily differentiate instruction in a way that challenges students at multiple achievement levels or allows them to more easily categorize students with less information. Teachers' experience at the time of the intervention was gleaned from the district's administrative records.[25] Consistent with the dial-setting model and the results on special education students, treatment effects are more negative and pronounced for less experienced teachers, particularly less experienced teachers who teach math.

Given that students (may) have different teachers in math and reading, results based on teacher characteristics are presented separately for math and reading exams. Students who have math teachers with less than three years of experience in HISD have treatment effects of -0.17σ (0.03) on math high-stakes scores whereas students with more experienced teachers have treatment effects of 0.01σ (0.04).[26] Effects on low-stakes math scores are similar. There is no difference in effect size for students with experienced or inexperienced reading teachers. The p-value for the treatment coefficient being different for each group is 0.000 in math on both high- and low-stakes scores and 0.937 and 0.197 in reading on high- and low-stakes scores, respectively.

In a similar spirit, I estimate the impact of treatment on students who have teachers of the same race as themselves versus students who have teachers of a different race. In math, effects are more negative for students whose teacher is of a different race than their own. Students with teachers of the same race have treatment effects of -0.024σ (0.04) on high-stakes test scores whereas students with teachers of a difference race have treatment effects of -0.11σ (0.04). The p-value on the difference is 0.019. The pattern on low-stakes

[25] Teachers' years of experience in HISD is the most accurate measure of teaching experience reported in administrative data. The correlation coefficient between age and HISD experience is 0.67 and the relationship is linear throughout the age distribution; these results are robust to splitting the sample on teachers' ages instead of HISD experience. The correlation between teachers' years of experience in HISD and teachers' self-reported years of experience on the teacher survey is 0.75 for teachers in the experimental sample.

[26] These results are robust to splitting the sample on teacher experience at any integer between one and six years of experience; the cutoff point was chosen to balance the sample between experienced and inexperienced teachers. The relatively high percent of inexperienced teachers (40 percent of teachers in HISD have fewer than 3 years of experience, and 20 percent are new teachers) may seem surprisingly high. However, Texas is a Right-to-Work state and HISD in particular has a culture of high rates of teacher turnover. There is no effect of treatment on teacher turnover, and the average years of teacher experience is balanced between treatment, control, and non-experimental schools.

scores is similar although the difference is only not significant. There is no difference in reading.

In addition to potential heterogeneity due to differences by subgroup in the tradeoff between the costs and benefits of specialization, there are potentially interesting differences for groups with varied exposure to teacher specialization.

Recall, there was greater potential for specialization in grade-language cells with English-language instruction relative to the cells with Spanish-language instruction in grades 1-3 (for details see Section III). Of the 52 Spanish-instruction cells in treatment schools, 25 percent had only one teacher in 2013-14, meaning that there was no margin for specialization. In each of the 71 English-instruction cells, there were at least two teachers. I focus on low-stakes scores given that high-stakes exams are only administered in grades 3-5. Limiting the analysis to grades 1-3, the treatment effect on low-stakes test scores for students who are *not* designated LEP is -0.17σ (0.05) versus -0.04σ (0.09) for students who are designated LEP. The p-value on the difference is 0.149. This is consistent with the fact that there was greater potential for specialization in English-instruction cells in the early grades. In grades 4 and 5, where teachers could specialize regardless of the language of instruction due to the structure of the transitional bilingual program, there is a -0.09σ (0.07) treatment effect per year for LEP students and a -0.10σ (0.04) treatment effect per year for non-LEP students on low-stakes exams.

Finally, there is potential heterogeneity in effects for students of different grade levels. It is possible that specialization is more effective in older elementary grades – where transition costs may be lower, and students are closer to the grades in which schools across the globe traditionally begin specializing their teachers. Again, I focus on the low-stakes results for this analysis, since only students in grades 3-5 take the high-stakes exams. Results are similar using high-stakes test scores. The ITT effect of instructing elementary schools to specialize is consistently negative and of similar magnitude for students in grades 1-4. The point estimates are slightly less negative for students in the 5th grade. Only the effects in grades 3 and 4 are statistically significant.

Furthermore, Table 6 demonstrates that there are important differences in the actual increase in the percent of potential specialization utilized in different grades. In grades 1 and 2, assignment to treatment increases the percent of specialization utilized by 36 percentage

points in year one (control mean = 0.23 and 0.30, respectively) and by 22 and 12 percentage points, respectively, in year two (control mean = 0.32 and 0.41, respectively). In grades 3-5, assignment to treatment only increases the percent of potential specialization utilized by 7-16 percentage points in year one (control mean = 0.52-0.66) and 5-11 percentage points in year two (control mean = 0.56-0.74).

Appendix Table 4 presents grade-specific results that scale the ITT effect by the increase in the percent of potential specialization utilized that was induced by random assignment to treatment. There is substantial heterogeneity in these LATE estimates by grade. One explanation that is consistent with these results is that the quality of additional specialization induced by assignment to treatment was much lower quality than the quality of marginal specialization in lower grades. This seems plausible, given that higher grades were already relatively highly specialized compared to the lower grades.

VI. Robustness Checks

In this subsection I explore the robustness of these results under potential threats to the interpretation of the data.

Attrition and Bounding

A concern for estimation is that I only include students for which I have post-treatment test scores. If students in treatment schools and students in control schools have different rates of selection into this sample, these results may be biased. Appendix Table 8 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 high-stakes 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. There is no treatment effect on attrition due to missing test scores. To see whether attrition affects these estimates, I provide Lee (2009) bounds on the main results in Table 3, 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 this case, however, because the attrition rates are quite similar between treatment and control, qualitatively the treatment effects remain unchanged.[27]

School-Level Heterogeneity

In the main analysis I use matched-pair fixed effects and school-year clustered standard errors. Abadie and Imbens (2011) show that the inclusion of matched-pair fixed effects should, in general, yield consistent standard errors with simple heteroskedasticity-robust standard errors. Yet, this may not correct for school-level heterogeneity in finite samples. This heterogeneity is uncorrelated with treatment due to random assignment, but could affect inference (Moulton 1986, 1990). Therefore, all of the main results presented above are estimated with more conservative clustered standard errors. I further address this issue in two ways: first, I estimate school-level regressions of the impact of treatment, and second, I calculate exact p-values via a nonparametric permutation test (Fisher 1935, Rosenbaum 1988).

I estimate unweighted school-level regressions of the impact on test scores in each treatment year, controlling for matched-pair fixed effects. The main results remain negative but are less precise (reported in Table 3). Appendix Table 9 presents school-level results for test scores disaggregated by subject and for behavior and attendance as well as teacher retention. All results are of the same sign as estimates using individual data but are statistically insignificant.

Next, I conduct a nonparametric permutation test as in Rosenbaum (1988). The sample is re-randomized 10,000 times between matched pairs at the school-level, like the original randomization. I re-calculate the ITT regressions with the new, synthetic treatment assignments and record the new treatment effects. The exact p-value is the proportion of

[27] Consistent with Lee (2009), the covariates used in the bounded results are the same covariates as those included when predicting attrition. Also, I use the same covariates to predict attrition as I do in the main analysis, so comparison between the main and bounded results is straightforward.

simulated treatment effects that are larger than the actual observed treatment effect (in absolute value).

Table 3 includes the exact p-values for the main results calculated via the permutation tests. Appendix Figure 1 plots the actual observed ITT treatment effect against the distribution of simulated treatment effects for various outcomes and subgroups. The negative effect on high- and low-stakes scores in the first year of treatment remains marginally significant. From the subgroup analysis, the large negative effects on high- and low-stakes scores for students with special needs remain highly significant; the effect on high-stakes math scores for students with inexperienced math teachers remains marginally significant. Furthermore, the highly negative slope on the dose-response graph remains highly significant when calculated via a permutation test with the full sample of students (grades 1-5) for all grade-language cells or for the grade-language cells with the most potential to specialize.

Together, these results confirm the basic facts described throughout; teacher specialization, if anything, lowers student achievement, particularly among the vulnerable subgroups (in a dial-setting, not economic, sense) described above.

Alternative Subjects

Recall that specialization also impacted science and social studies classes. In some schools, the teacher specialized to teach math also taught science and the teacher specialized to reading also taught social studies; in schools with more teachers, one teacher taught math, one taught reading, and a third taught science and social studies. In either scenario one would expect there to be an effect of specialization on science and social studies scores as well as math and reading.

Appendix Table 10 examines the effect of treatment on students' performance on low-stakes Stanford 10/ITBS science and social studies tests.[28] Panel A displays results for grades 1 though 5. Columns (1)-(3) present ITT estimates of the effects on science and social studies scores, which were negative and significant in both years of the experiment: treatment *lowered* science and social studies achievement *each* by 0.07σ (0.02) per year.

[28] In 2014, the Stanford 10 science and social studies exams were administered to students in grades 3-5. In 2015, the ITBS science and social studies exams were administered to students in grades 1-5.

Consistent with the main results in math and reading, teacher specialization seems to decrease student test scores.

Multiple Hypothesis Testing

I have run many regressions with various outcomes in differing subsamples to measure treatment effects. A concern is that I am simply detecting false positives due to multiple hypothesis-testing. Appendix Table 11 presents results that control for the family-wise error rate, which is defined as the probability of making one or more false discoveries – known as type I errors – when performing multiple hypothesis tests, using the (conservative) Holm step-down method described in Romano, Shaikh and Wolf (2010).

Appendix Table 11 confirms the robustness of our main findings. All main effects and the main subsample effects that are highly significant in the main analysis remain significant after accounting for multiple hypothesis testing.

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

The experiment analyzed 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.[29]

[29] Cook and Mansfield (2016) consider the effects of re-allocating high-school teachers to the subjects they are more effective at teaching – a case in which there are likely few additional costs, since teachers are already specialized and simply being shuffled around to teach different subjects. They find that having a teacher in a given subject whose subject-specific skill in that subject is 1σ greater than his or her average across all subjects is associated with a 0.06σ increase in student test scores.

The increase in student achievement caused due to teacher sorting on comparative advantage can be indirectly computed. A measure of how much a teacher has to gain from specialization is the difference between a teacher's TVA is in the subject he is sorted to teach in a treatment school and his average TVA in all subjects he used to teach before. I also calculate treatment effects for each matched pair using individual-level data and assign a teacher the matched-pair treatment effect of the students assigned to their classroom in a given subject. Figure 4 plots the relationship between the treatment effect on a teacher's students and how much the teacher stood to gain from specialization.[30] Students with teachers who stand to gain the least from sorting have the largest *decreases* in student achievement in both math and reading. Teachers with the most to gain from sorting may actually experience gains from specialization. In year one, students whose math [reading] teacher's math [reading] TVA is 1σ greater than his or her average math and reading TVA have high-stakes treatment effects that are 0.15σ [0.05σ] higher than students whose math [reading] teacher's math [reading] TVA is equal to his or her average math and reading TVA (p-value 0.003 [0.053]). However, the relationship is negative and insignificant for both math and reading teachers in year two. Results are similar using low-stakes test scores.

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.

Table 7 reports treatment effects from the ITT specification 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.

[30] A similar pattern emerges if the relationship is plotted at the school level, with the matched-pair treatment effect on the y-axis and school average of teacher difference in TVA from sorted subject to the mean of all subjects taught in the previous year on the x-axis. See Appendix Figure 1, Panel A.

There is some suggestive evidence that inefficient dial-setting may explain a portion of the results. Treatment teachers are 4.5 (2.7) percentage points less likely to report they "know" their students (control mean = 81.8%) and 8.9 (2.5) percentage points less likely to report providing them with individual attention (control mean = 62.0%). 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.

Additionally, there was a negative effect of treatment on teachers' self-reported job satisfaction and performance, relative to the previous year. Treatment teachers were 17.0 (4.7) percentage points less likely to report an above-median increase in job satisfaction (control mean 48.1%) and 15.9 (5.2) percentage points less likely to report an above-median increase in job performance (control mean 47.1).

Given that the negative effect on student achievement was larger for students with inexperienced teachers, the remaining columns of Table 7 split the survey results by teacher experience level. Treatment teachers with less than three years of experience are 11.4 (4.0) percentage points less likely to report that they "know" their students (control mean 90.7%) and 19.1 (6.4) percentage points less likely to report providing them with individual attention (control mean 59.2%). Both results are statistically significant. Conversely, there is no effect on either "knowing" students or providing them with individual attention for more experienced teachers. Both experienced and inexperienced teachers have large and significant negative effects on job satisfaction and performance.

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. However, these results are largely robust to a standard bounding procedure that takes this differential response rate into account (See Appendix Table 12).

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.

This paper is the first field experiment – and I hope not the last – on teacher specialization in elementary schools. Future research might focus on whether specialization is differentially effective for schools with different baseline rates of specialization, and attempt to clarify the mechanisms behind the effects of specializing teachers in elementary schools. That said, these results provide a cautionary tale about the potential productivity benefits of the division of labor when applied to early 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.

Condie, S., Lefgren, L. & Sims, D. (2014). Teacher heterogeneity, value-added and education policy. *Economics of Education Review*. 40: 76-92.

Cook, J. B. and Mansfield, R. K. (2016). Task-specific experience and task-specific talent: Decomposing the productivity of high school teachers. *Journal of Public Economics* 140: 51-72.

Delpit, L. (2006). Other People's Children: Cultural Conflict in the Classroom. 1R edition. New York: The New Press.

Fryer, R. G. (2014) Injecting charter school best practices into traditional public schools: Evidence from field experiments. *Quarterly Journal of Economics*. 129(3): 1355-1407.

Gewertz, C. (2014). 'Platooning' on the rise in early grades. *Education Week,* 33(21): 1,16-17.

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.

Jacob, B. A. & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics,* 26(1): 101-136

Jacob, B. A. & Rockoff, J.E. (2011). Organizing schools to improve student achievement: start times, grade configurations, and teacher assignments. Brookings Institute: *The Hamilton Project* Discussion Paper 2011-08.

Jovanovic, B., and Rousseau, P. L.. (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, D. 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, B. W. (1989). The Model T Ford Encyclopedia, 1909-1927: A comprehensive guide to the evolution and changes of the major components 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.

Rockoff, J.E., Staiger, D. O., Kane, T. J., and Taylor, E. S. (2010) "Information and employee evaluation: Evidence from a randomized intervention in public schools" *NBER Working Paper No. 16240*.

Romano, J. P., Shaikh A. M., and Wolf, M.. 2010. "Hypothesis testing in econometrics." *Annual Review of Economics*, 2: 75-104.

Rosenbaum, P. 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.

U.S. Department of Education, National Center for Education Statistics, Schools and Staffing Survey, Public Teachers Data File 2011-12.

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: Pre-Treatment Summary Statistics

	Non-Exp Mean	Exp Mean	*p-value*	Control Mean	Treatment Mean	*p-value*
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: School Characteristics						
Percent female	0.488	0.497	0.194	0.493	0.500	0.499
Percent Black	0.227	0.377	0.008	0.386	0.368	0.858
Percent Hispanic	0.632	0.598	0.556	0.591	0.604	0.902
Percent White	0.086	0.014	0.000	0.010	0.019	0.398
Percent Asian	0.043	0.006	0.000	0.008	0.005	0.597
Percent other race	0.012	0.005	0.000	0.005	0.005	0.902
Percent limited English proficient	0.438	0.403	0.390	0.389	0.418	0.670
Percent receiving special education services	0.047	0.049	0.620	0.047	0.050	0.623
Percent gifted and talented	0.190	0.130	0.000	0.129	0.131	0.931
Percent economically disadvantaged	0.806	0.937	0.000	0.937	0.937	0.993
Percent committed behavioral offense 12-13	0.042	0.054	0.247	0.048	0.060	0.334
Mean attendance rate 12-13	97.020	96.689	0.039	96.816	96.561	0.333
Mean STAAR Math Score 12-13 (σ units)	0.044	-0.308	0.000	-0.294	-0.323	0.752
Mean STAAR Reading Score 12-13 (σ)	0.013	-0.221	0.000	-0.214	-0.228	0.846
School's degree of specialization 12-13 (Actual)	0.289	0.273	0.360	0.275	0.271	0.894
School's degree of specialization 12-13 (Ideal)	0.803	0.815	0.411	0.808	0.822	0.556
School's degree of specialization 12-13 (Actual/Ideal)	0.378	0.349	0.308	0.355	0.343	0.805
Number of Schools	127	46		23	23	
p-value from joint F-test			*0.0000*			*0.0632*
Panel B: Teacher Characteristics						
Teacher experience	7.746	8.084	0.493	8.563	7.669	0.293
Math TVA in 12-13 (σ)	0.163	-0.071	0.005	0.010	-0.142	0.292
Reading TVA in 12-13 (σ)	0.025	-0.203	0.000	-0.119	-0.278	0.144
Teacher does not teach both math and reading in 12-13	0.338	0.309	0.458	0.326	0.294	0.616
Teacher teaches 3 or fewer subjects in 12-13	0.381	0.367	0.750	0.342	0.389	0.541
Teacher teaches 2 or fewer subjects in 12-13	0.294	0.258	0.301	0.272	0.245	0.630
Number of subjects taught in 12-13	3.236	3.293	0.523	3.294	3.291	0.982
Number of grades taught in 12-13	1.056	1.032	0.114	1.035	1.029	0.763
Number of student contacts in 12-13	35.213	34.840	0.842	34.788	34.888	0.973
Number of Teachers	2705	977		456	521	
p-value from joint F-test			*0.0000*			*0.0002*
Panel C: Student Characteristics						
Female	0.493	0.497	0.396	0.491	0.501	0.245
Black	0.182	0.331	0.004	0.340	0.323	0.862
Hispanic	0.662	0.646	0.770	0.635	0.655	0.847
White	0.098	0.013	0.000	0.011	0.014	0.621
Asian	0.045	0.006	0.000	0.008	0.004	0.431
Other Race	0.013	0.005	0.000	0.005	0.005	0.912
Limited English Proficient	0.465	0.442	0.577	0.424	0.458	0.595
Special Education Services	0.046	0.047	0.735	0.047	0.047	0.965
Gifted and Talented	0.206	0.137	0.000	0.136	0.138	0.911
Economically Disadvantaged	0.786	0.937	0.000	0.933	0.940	0.593
Commit Behavioral Offense 12-13	0.031	0.049	0.005	0.042	0.055	0.269
Attendance Rate 12-13	97.147	96.800	0.021	96.933	96.683	0.353
STAAR Math Score 12-13 (σ)	0.131	-0.272	0.000	-0.251	-0.290	0.665
STAAR Reading Score 12-13 (σ)	0.088	-0.191	0.000	-0.181	-0.198	0.824
Number of Students	54989	18701		8790	9911	
p-value from joint F-test			*0.0000*			*0.7211*

Notes: This table reports school-, teacher-, and student-level pre-treatment summary statistics. Students are only included in the sample if they have at least one valid high- or low-stakes test score outcome variable in 2014-15 and are enrolled in grades 1-5. Column (1) reports the mean of the non-experimental group. Column (2) reports the mean of the experimental group. Column (3) reports the p-value on the null hypothesis of equal means in the experimental and non-experimental groups. Columns (4)-(6) report similar values for the treatment versus control group. The tests in Columns (3) and (6) use heteroskedasticity-robust standard errors in Panel A and school-clustered standard errors in Panel B and Panel C. All demographic and test score measures are culled from administrative data collected pre-treatment. See the Online Appendix for details on variable construction. Student test scores and teacher effect measures are standardized to have a mean of zero and standard deviation one over the district sample by grade and by subject, respectively. Measures of school specialization are calculated from administrative data – for details see the Online Appendix.

Table 2A: Proof of Treatment Effect on Teacher Specialization, All Teachers

	2013-2014				2014-2015	
	Survey Data		Admin Data		Admin Data	
	Control Mean	ITT	Control Mean	ITT	Control Mean	ITT
	(1)	(2)	(3)	(4)	(5)	(6)
Does not teach math and reading	0.514	0.362***	0.476	0.372***	0.549	0.267***
		(0.043)		(0.039)		(0.057)
Observations		666		977		982
Teaches 3 or fewer subjects	0.588	0.332***	0.507	0.428***	0.586	0.278***
		(0.040)		(0.047)		(0.061)
Observations		666		977		982
Teaches 2 or fewer subjects	0.500	0.259***	0.406	0.317***	0.503	0.195***
		(0.037)		(0.036)		(0.059)
Observations		666		977		982
Self-report departmentalized	0.560	0.360***	—	—	—	—
		(0.049)				
Observations		649				
Number of subjects taught	2.672	-0.770***	2.939	-0.917***	2.712	-0.561***
		(0.100)		(0.086)		(0.139)
Observations		666		977		982
Number of grades taught	1.689	-0.176	1.024	0.053**	1.022	0.017
		(0.129)		(0.022)		(0.011)
Observations		693		977		982
Total # of student contacts	—	—	37.724	11.240***	42.418	6.682***
				(1.785)		(2.332)
Observations				977		982
% of potential degree of specialization	—	—	0.427	0.258***	0.528	0.133***
				(0.028)		(0.047)
Observations				977		982

Notes: This table presents estimates of the effect of being enrolled in a treatment school on the degree of specialization of each teacher as measured by administrative data and survey responses. The sample includes all teachers employed in HISD in each year with valid measures of specialization. The measure in the final row of the table is an author-constructed continuous measure of the degree of specialization of each teacher based on the number of subjects taught and the rightness of fit of the teacher to the subjects that she or he is teaching, conditional on the other teachers teaching in the same grade-language cell, divided by the degree of specialization of that teacher under ideal specialization. All presented survey responses except the number of subjects or grades taught or the average number of students per subject per grade have been made into indicator variables. Each regression controls for matched pair fixed effects. See the Online Appendix for details on the construction of the variables used in the table. Columns (2), (4), and (6) contain coefficients on an indicator for working at a treatment school. Columns (1)-(4) display estimates from the first year of treatment (survey and administrative data) and Columns (5)-(6) display estimates from the second year of treatment (administrative data only). Teachers are assigned to treatment if they taught in a treatment school in each year. All standard errors, located in parentheses, are clustered by school. *, **, and *** denote significance at the 90, 95, and 99 percent confidence levels, respectively.

Table 2B: Proof of Treatment Effect on Teacher Specialization, Teachers Who Were Not Specialized in Pre-Treatment Year

	2013-2014				2014-2015	
	Survey Data		Admin Data		Admin Data	
	Control Mean	ITT	Control Mean	ITT	Control Mean	ITT
	(1)	(2)	(3)	(4)	(5)	(6)
Does not teach math and reading	0.395	0.501***	0.363	0.463***	0.491	0.308***
		(0.050)		(0.044)		(0.060)
Observations		477		772		847
Teaches 3 or fewer subjects	0.463	0.451***	0.399	0.531***	0.534	0.321***
		(0.049)		(0.054)		(0.065)
Observations		477		772		847
Teaches 2 or fewer subjects	0.361	0.400***	0.306	0.416***	0.445	0.244***
		(0.043)		(0.038)		(0.061)
Observations		477		772		847
Self-report departmentalized	0.450	0.484***	—	—	—	—
		(0.062)				
Observations		472				
Number of subjects taught	3.005	-1.108***	3.195	-1.163***	2.855	-0.661***
		(0.106)		(0.092)		(0.140)
Observations		477		772		847
Number of grades taught	1.346	-0.116	1.028	0.050*	1.025	0.010
		(0.102)		(0.026)		(0.011)
Observations		482		772		847
Total # of student contacts	—	—	33.564	14.054***	40.265	7.249***
				(1.770)		(2.233)
Observations				772		847
% of potential degree of specialization	—	—	0.351	0.336***	0.485	0.165***
				(0.028)		(0.048)
Observations				772		847

Notes: This table presents estimates of the effect of being enrolled in a treatment school on the degree of specialization of each teacher that was not already specialized in the year previous to treatment. as measured by administrative data and survey responses. The measure in the final row of the table is an author-constructed continuous measure of the degree of specialization of each teacher based on the number of subjects taught and the rightness of fit of the teacher to the subjects that she or he is teaching, conditional on the other teachers teaching in the same grade-language cell, divided by the degree of specialization of that teacher under ideal specialization. All presented survey responses except the number of subjects or grades taught or the average number of students per subject per grade have been made into indicator variables. Each regression controls for matched pair fixed effects. See the Online Appendix for details on the construction of the variables used in the table. Columns (2), (4), and (6) contain coefficients on an indicator for working at a treatment school. Columns (1)-(4) display estimates from the first year of treatment (survey and administrative data) and Columns (5)-(6) display estimates from the second year of treatment (administrative data only). Teachers are assigned to treatment if they taught in a treatment school in each year. All standard errors, located in parentheses, are clustered by school. *, **, and *** denote significance at the 90, 95, and 99 percent confidence levels, respectively.

Table 3: The Effect of Treatment on Student Test Scores (ITT)

	High-Stakes			Low-Stakes		
	2014	2015	Pooled	2014	2015	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment Effect (σ units)	-0.122**	-0.089*	-0.106***	-0.109**	-0.090	-0.100**
	(0.046)	(0.051)	(0.038)	(0.047)	(0.065)	(0.042)
Control Mean (σ)	-0.239	-0.289	-0.264	-0.232	-0.264	-0.247
Observations	10,462	10,360	20,822	18,618	16,849	35,467
Robustness Checks						
Lee Bounds (σ)	-0.160***	-0.136***	-0.146***	-0.133***	-0.097	-0.111***
	(0.046)	(0.050)	(0.037)	(0.044)	(0.065)	(0.041)
School-Level Regressions (σ)	-0.141*	-0.070	-0.106	-0.123	-0.076	-0.103
	(0.077)	(0.072)	(0.068)	(0.075)	(0.084)	(0.076)
P-value, Permutation Test	[0.079]	[0.242]	[0.119]	[0.119]	[0.348]	[0.201]

Notes: This table presents estimates of the effect of being enrolled in a treatment school on high- and low-stakes test scores. High-stakes test scores are summed math and reading STAAR scores and low-stakes scores are summed math and reading Stanford 10 scores (in year 1) or ITBS scores (in year 2). Here treatment is defined as attending a treatment school as the last school in 2012-13. The sample is restricted each year to those students who are attending grades 3 through 5 (for high-stakes exams) and grades 1 through 5 (for low-stakes exams) and have both valid math and reading test scores. All columns report Intent-to-Treat (ITT) estimates. Columns (1) and (4) use 2013-2014 scores as the outcome variable. Columns (2) and (5) use 2014-2015 scores as the outcome variable. Columns (3) and (6) use scores from both 2013-2014 and 2014-2015 as the outcome variable. The dependent variable in all specifications is the sum of standardized math and reading test scores, (standardized across the district 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 indicators for taking a Spanish baseline test. All specifications have grade-by-year and matched-pair fixed effects. Standard errors, reported in parentheses, are clustered at the school-year level. 90%, 95%, and 99% confidence levels are indicated by *, **, and ***, respectively. The following robustness checks are provided in the final rows of the table: a bounded estimate that accounts for differential attrition rates between treatment and control using the methods described in Lee (2009); school-level regression results (control variables are matched pair fixed effects only; N=46 in all regressions); and the exact p-value for the treatment coefficient in the main results calculated via a permutation test (Fisher 1935, Rosenbaum 1988).

Table 4: The Effect of Treatment on Student Test Scores (2SLS - Student Attendance)

	ITT			2SLS (Ever)			2SLS (Years)		
	2014	2015	Pooled	2014	2015	Pooled	2014	2015	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
High-Stakes	-0.122**	-0.089*	-0.106***	-0.137***	-0.103*	-0.121***	-0.147***	-0.057*	-0.089***
	(0.046)	(0.051)	(0.038)	(0.050)	(0.059)	(0.043)	(0.054)	(0.032)	(0.032)
Observations	10,462	10,360	20,822	10,462	10,360	20,822	10,462	10,360	20,822
First Stage Coefficient				0.896***	0.858***	0.877***	0.833***	1.552***	1.191***
				(0.009)	(0.012)	(0.008)	(0.009)	(0.025)	(0.042)
Low-Stakes	-0.109**	-0.090	-0.100**	-0.122**	-0.105	-0.115**	-0.132**	-0.059	-0.086**
	(0.047)	(0.065)	(0.042)	(0.052)	(0.075)	(0.047)	(0.056)	(0.042)	(0.036)
Observations	18,618	16,849	35,467	18,618	16,849	35,467	18,618	16,849	35,467
First Stage Coefficient				0.889***	0.854***	0.872***	0.823***	1.533***	1.161***
				(0.009)	(0.012)	(0.008)	(0.010)	(0.026)	(0.042)

Notes: This table presents estimates of being enrolled in or attending a treatment school on high- and low-stakes test scores. High-stakes test scores are summed math and reading STAAR scores and low-stakes scores are summed math and reading Stanford 10 scores (in year 1) or ITBS scores (in year 2). Here treatment is defined as attending a treatment school as the last school in 2012-13. The sample is restricted each year to those students who are attending grades 3 through 5 (for high-stakes exams) and grades 1 through 5 (for low-stakes exams) and have both valid math and reading test scores.Here treatment is defined as attending a treatment school as the last school in 2012-13. 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 during years of treatment. 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 scores as the outcome variable. Columns (2), (5), and (8) use 2014-2015 scores 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 the sum of standardized math and reading test scores, (standardized across the district 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 indicators for taking a Spanish baseline test. All specifications have grade, year, and matched-pair fixed effects. The final row provides the first stage coefficient of instrumenting the 2SLS *Ever* or *Years* variable with ITT treatment assignment. This number can be used to scale the ITT estimate into other estimates. Standard errors, reported in parentheses, are clustered at the school-year level. 90%, 95%, and 99% confidence levels are indicated by *, **, and ***, respectively.

Table 5: Sensitivity Analysis or Extension of the Basic Model

	High-Stakes	*p-value*	Obs	Low-Stakes	*p-value*	Obs
	(1)	(2)	(3)	(4)	(5)	(6)
Full Sample (pooled)	-0.106***		20,822	-0.100**		35,467
	(0.038)			(0.042)		
Panel A: Demographics						
Special Education: Yes	-0.296***	0.037	688	-0.219***	0.088	1,424
	(0.090)			(0.068)		
Special Education: No	-0.101**		19,346	-0.096**		32,633
	(0.039)			(0.043)		
Gifted: Yes	-0.040	0.041	3,603	-0.068	0.278	4,220
	(0.062)			(0.054)		
Gifted: No	-0.133***		16,431	-0.114**		29,837
	(0.038)			(0.044)		
LEP: Yes	-0.079	0.621	8,856	-0.042	0.164	15,406
	(0.071)			(0.074)		
LEP: No	-0.114***		11,178	-0.143***		18,651
	(0.033)			(0.036)		
LEP: Yes (Grade 1-3)	-0.109	0.794	3,166	-0.040	0.149	9,570
[High Stakes - Gr. 3 Only]	(0.109)			(0.093)		
LEP: No (Grade 1-3)	-0.139**		3,736	-0.171***		11,064
[High Stakes - Gr. 3 Only]	(0.059)			(0.047)		
LEP: Yes (Grade 4-5)	-0.074	0.787	5,690	-0.093	0.936	5,836
	(0.066)			(0.071)		
LEP: No (Grade 4-5)	-0.093**		7,442	-0.099**		7,587
	(0.041)			(0.041)		
Panel B: Grade Levels						
Grade 1	—			-0.109	0.621	6,887
				(0.073)		
Grade 2	—			-0.088		7,325
				(0.074)		
Grade 3	-0.133**	0.586	7,189	-0.136**		7,312
	(0.060)			(0.061)		
Grade 4	-0.124**		7,058	-0.149**		7,222
	(0.056)			(0.058)		
Grade 5	-0.068		6,575	-0.060		6,721
	(0.050)			(0.046)		
Panel C: Teacher Characteristics						
Math:						
Teacher experience $<$ 3 years	-0.173***		6,435	-0.106***		9,762
	(0.030)			(0.028)		
Teacher experience $\geq$ 3 years	0.003	0.000	10,401	-0.002	0.004	18,816
	(0.035)			(0.033)		
Reading:						
Teacher experience $<$ 3 years	-0.057		6,419	-0.106***		9,098
	(0.040)			(0.035)		
Teacher experience $\geq$ 3 years	-0.060**	0.937	10,330	-0.055*	0.197	19,445
	(0.025)			(0.029)		
Math:						
Above-Median Potential Gains Spec.	-0.055		3,010	-0.049		7,835
	(0.057)			(0.050)		
Below-Median Potential Gains Spec.	-0.043		6,941	0.002		9,298
	(0.034)			(0.039)		

Missing Potential Gains Spec.	-0.098***	0.344	7,563	-0.072**	0.201	12,610
	(0.029)			(0.031)		
Reading:						
Above-Median Potential Gains Spec.	-0.020		2,999	-0.092**		5,615
	(0.047)			(0.041)		
Below-Median Potential Gains Spec.	-0.093***		7,370	-0.063*		12,401
	(0.035)			(0.034)		
Missing Potential Gains Spec.	-0.057**	0.423	7,222	-0.077***	0.780	11,796
	(0.027)			(0.028)		
Math:						
Teacher-Student Same Race	-0.024		5,558	-0.034		10,363
	(0.040)			(0.039)		
Teacher-Student Different Race	-0.112***	0.019	3,643	-0.075**	0.228	5,899
	(0.037)			(0.031)		
Reading:						
Teacher-Student Same Race	-0.089**		5,506	-0.093***		10,438
	(0.037)			(0.031)		
Teacher-Student Different Race	-0.064**	0.440	3,892	-0.070**	0.509	5,936
	(0.025)			(0.033)		
Math:						
Has Student-Teacher Link	-0.068***	0.020	17,599	-0.044	0.078	29,921
	(0.026)			(0.027)		
Missing Student-Teacher Link	0.017		3,319	0.015		5,785
	(0.024)			(0.022)		
Reading:						
Has Student-Teacher Link	-0.056**	0.042	17,784	-0.070***	0.060	29,895
	(0.022)			(0.025)		
Missing Student-Teacher Link	0.005		3,256	-0.010		5,710
	(0.020)			(0.022)		

Notes: This table presents estimates of the effect of being enrolled in treatment school on high- and low-stakes test scores. In Panels A, B, and C, high-stakes test scores are summed math and reading STAAR scores and low-stakes scores are summed math and reading Stanford 10 scores (in year 1) or ITBS scores (in year 2). In Panel D (teacher characteristics), math and reading scores are reported separately since students (may) have different math and reading teachers. Teachers' potential gains from specialization is defined as the difference between their TVA in the subject they teach the student in and their average TVA in both math and reading. For details on all variables used to subset the sample, see the Online Appendix. Here treatment is defined as attending a treatment school as the last school in 2012-13. The sample is restricted each year to those students who are attending grades 3 through 5 (for high-stakes exams) and grades 1 through 5 (for low-stakes exams) and have both valid math and reading test scores. All columns report Intent-to-Treat (ITT) estimates and follow the pooled specification from Table 4, but teacher race is only available in 2013-14 and therefore those results present the year one specification. The dependent variable in all specifications are standardized math and reading test scores, standardized across the district to have a mean of zero and standard deviation one by grade and year. In Panels A, B, and C, the sum of math and reading is used. In Panel D, math and reading are reported separately. 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 indicators for taking a Spanish baseline test. All specifications have grade-by-year and matched-pair fixed effects. Standard errors, reported in parentheses, are clustered at the school-year level. 90%, 95%, and 99% confidence levels are indicated by *, **, and ***, respectively.

Table 6: The Effect of Treatment on the Degree of Specialization by Grade Level

	2014		2015	
	Control Mean	ITT	Control Mean	ITT
	(1)	(2)	(3)	(4)
Full Sample	0.442	0.219***	0.546	0.114**
		(0.027)		(0.047)
N		17,854		16,257
Grade 1	0.231	0.359***	0.339	0.205***
		(0.038)		(0.054)
N		3,643		2,778
Grade 2	0.298	0.358***	0.422	0.109*
		(0.052)		(0.061)
N		3,504		3,342
Grade 3	0.524	0.161***	0.625	0.109
		(0.048)		(0.084)
N		3,542		3,186
Grade 4	0.519	0.142***	0.564	0.105**
		(0.040)		(0.048)
N		3,765		3,537
Grade 5	0.664	0.073	0.738	0.049
		(0.057)		(0.056)
N		3,400		3,414

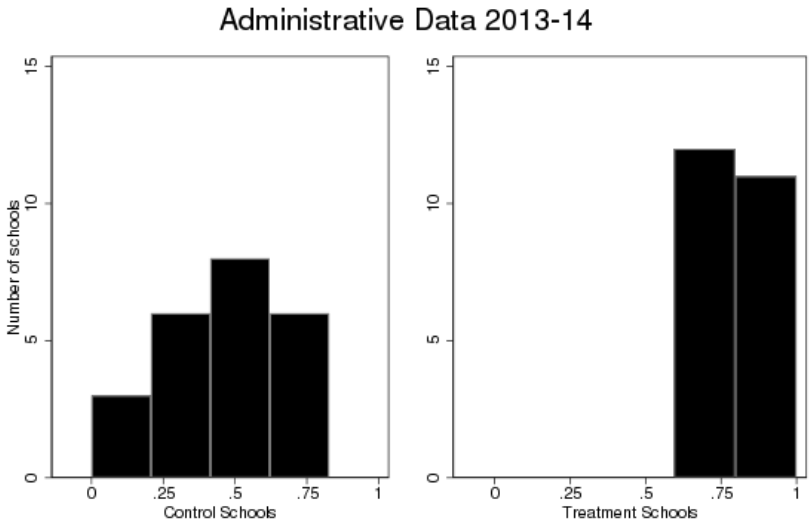
Notes: This table reports ITT results of the effect of treatment on the percent of potential specialization utilized for the full sample and for each grade. The dependent variable is the percent of potential specialization utilized in the grade-language cell that each student is assigned to. For details on the construction of this measure, see the Online Appendix. Columns (1) and (3) report the mean percent of potential specialization utilized of the control group. Columns (2) and (4) report the ITT effect of treatment on the percent of potential specialization utilized for each sample. The sample includes all students who have valid high- or low-stakes scores in each year who are enrolled in grades 1-5 in each year. All regressions adjust for the student-level demographics summarized in Table 1, student-level math and reading scores (3 years prior to the first year of treatment) and their squares, and indicators for taking a Spanish baseline test, as well as grade-by-year and matched pair fixed effects. Standard errors, reported in parentheses, are clustered at the school-year level. 90%, 95%, and 99% confidence levels are indicated by *, **, and ***, respectively.

Table 7: The Effect of Treatment on Survey Outcomes

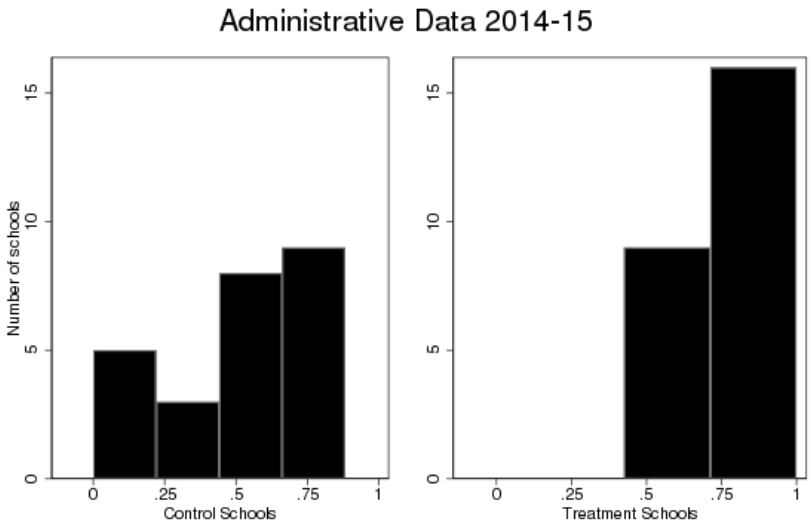
	Full Sample		Inexperienced Teachers		Experienced Teachers	
	Control Mean	ITT	Control Mean	ITT	Control Mean	ITT
	(1)	(2)	(3)	(4)	(5)	(6)
Know students	0.818	-0.045*	0.907	-0.114***	0.798	-0.013
		(0.026)		(0.040)		(0.030)
Observations		663		227		409
Gives individual attention	0.620	-0.089***	0.592	-0.191***	0.634	-0.048
		(0.025)		(0.064)		(0.037)
Observations		667		229		412
Rules enforced	0.678	-0.041	0.592	-0.084	0.716	-0.002
		(0.061)		(0.088)		(0.062)
Observations		670		229		414
Enthusiasm for teaching math	0.868	-0.026	0.861	-0.032	0.870	-0.017
		(0.025)		(0.034)		(0.030)
Observations		695		234		434
Enthusiasm for teaching reading	0.865	-0.013	0.881	-0.033	0.861	0.005
		(0.020)		(0.039)		(0.022)
Observations		698		234		436
Lesson differentiation	37.181	2.137	37.818	-0.999	38.109	2.638
		(1.942)		(3.419)		(2.636)
Observations		684		232		424
Above-median change in job satisfaction	0.481	-0.170***	0.396	-0.154**	0.542	-0.248***
		(0.047)		(0.062)		(0.059)
Observations		484		172		299
Above-median change in job performance	0.471	-0.159***	0.453	-0.216***	0.495	-0.175***
		(0.052)		(0.066)		(0.058)
Observations		484		172		299

Notes: This table presents estimates of the effect of being enrolled in a treatment school on teacher end-of-year survey responses. Each row presents effects on a different survey outcome. Columns (1)-(2) present results for all teachers and Columns (3)-(4) and (5)-(6) present results separately for teachers with fewer than 3 years of teaching experience and more than 3 years of experience, respectively (as reported in administrative data). 763 teachers returned the survey; response rates on the questions reported here range from 64 to 92 percent. All presented survey responses except lesson differentiation have been made into indicator variables. See Online Appendix B for details on the construction of the variables used in the table. All regressions include matched pair fixed effects. Standard errors, reported in parentheses, are clustered at the school level. 90%, 95%, and 99% confidence levels are indicated by *, **, and ***, respectively.

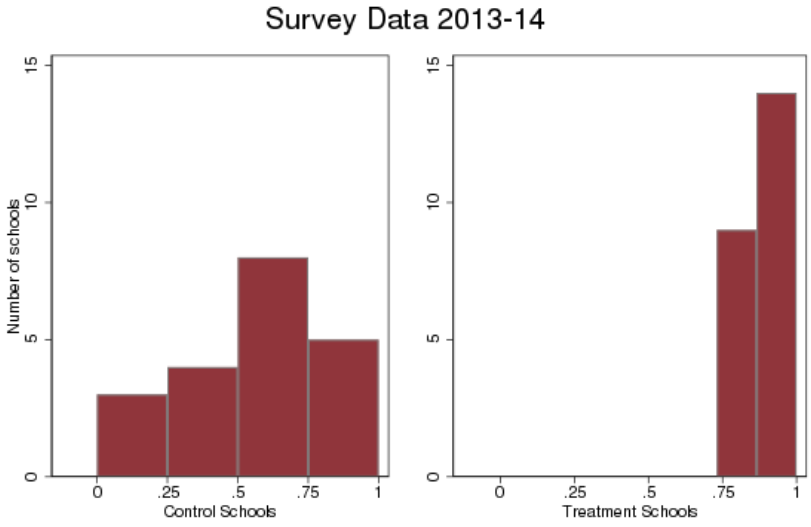
Figure 1: Distribution of the Percent of Teachers Specialized in Treatment and Control Schools



A specialized teacher is one who does not teach both math and reading in the student-teacher linkage file during the first year of treatment. Percent of teachers that are specialized is calculated over all teachers in the elementary school.

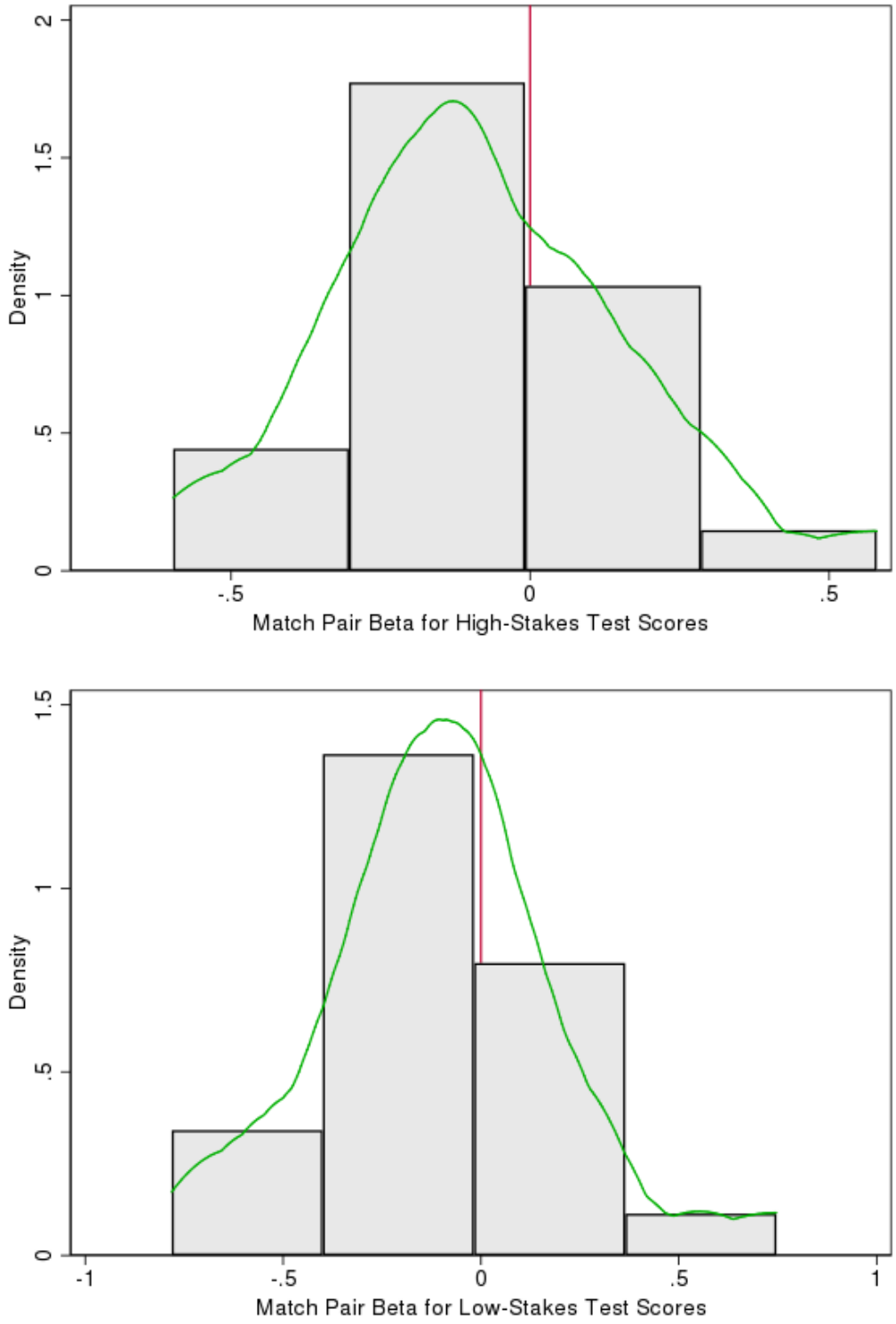


A specialized teacher is one who does not teach both math and reading in the student-teacher linkage file during the second year of treatment. Percent of teachers that are specialized is calculated over all teachers in the elementary school.



A specialized teacher is one who does not teach both math and reading as self-reported in the teacher survey at the end of one year of treatment. Percent of teachers that are specialized is calculated over all teachers that responded to the survey in the elementary school.

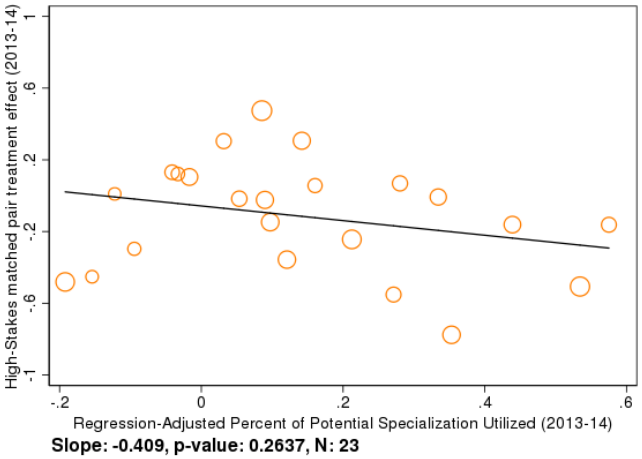
Figure 2: Matched Pair Specific Treatment Effects



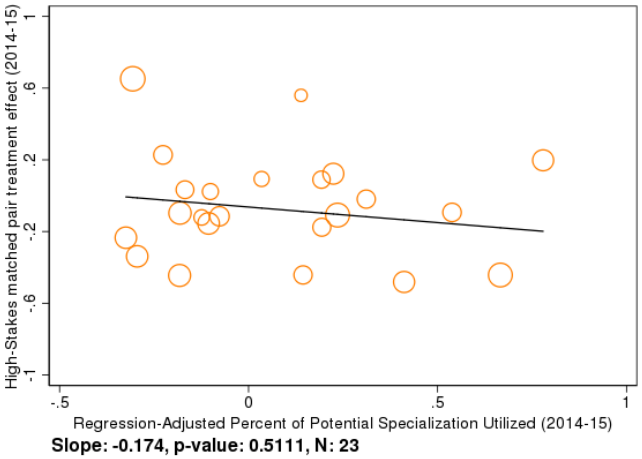
Notes: These figures plot a kernal density curve for matched pair treatment coefficients pooled over both years of treatment. Treatment coefficients are obtained by regressing the sum of math and reading high- or low-stakes test scores on a treatment indicator, student demographics, baseline test scores and grade-by-year fixed effects for each matched pair. The figures also display a red vertical line at 0 for comparison

Figure 3: Dose Response

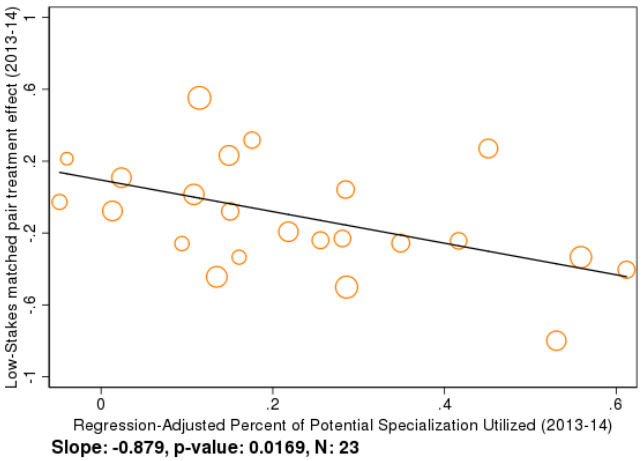
Panel A: Full Sample



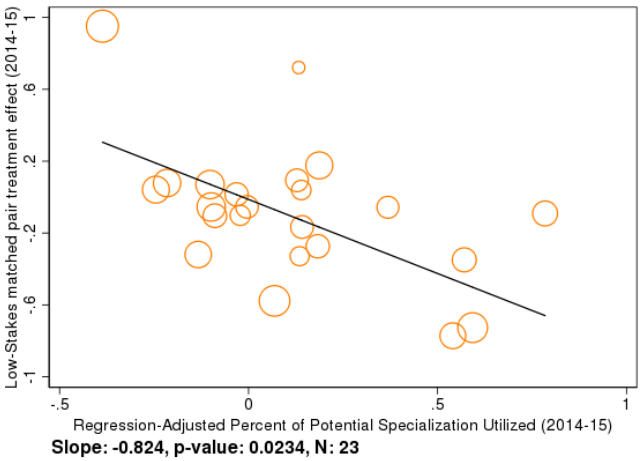
Slope: -0.409, p-value: 0.2637, N: 23



Slope: -0.174, p-value: 0.5111, N: 23

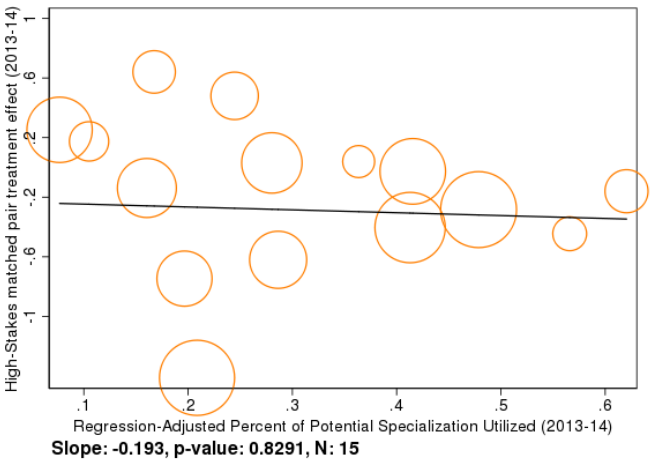


Slope: -0.879, p-value: 0.0169, N: 23

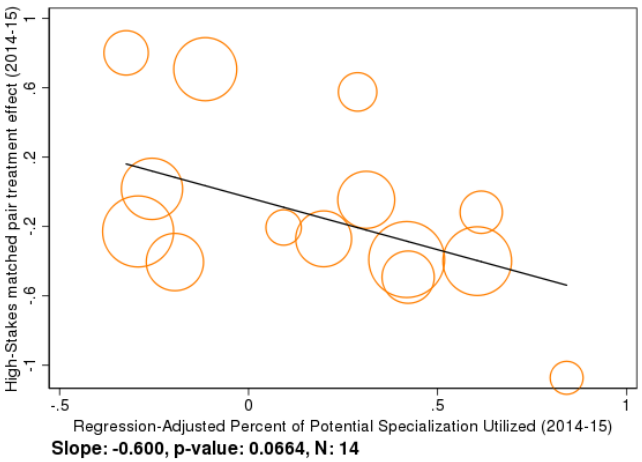


Slope: -0.824, p-value: 0.0234, N: 23

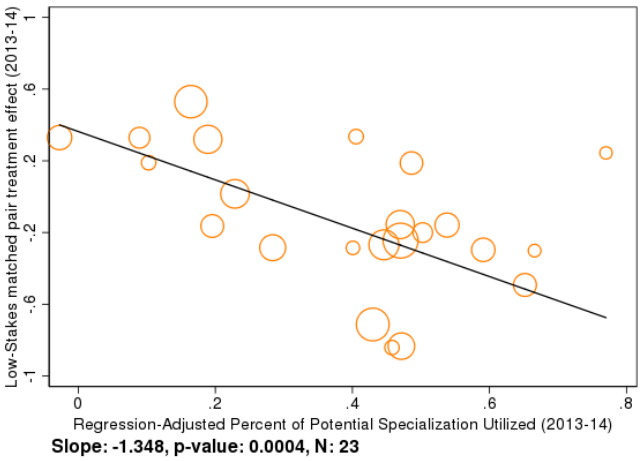
Panel B: Unspecialized Grade-Language Cells in Pre-Treatment Year



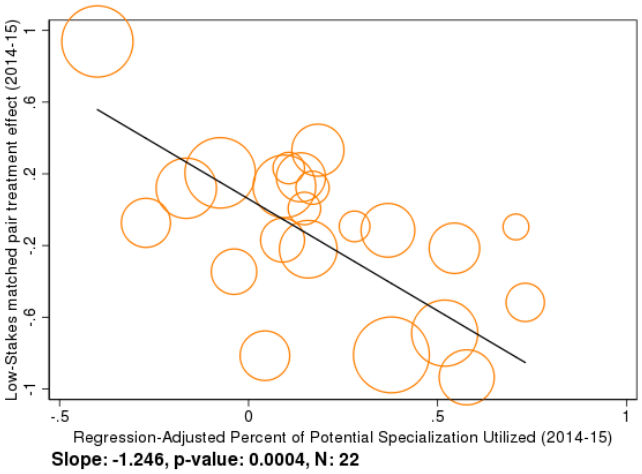
Slope: -0.193, p-value: 0.8291, N: 15



Slope: -0.600, p-value: 0.0664, N: 14

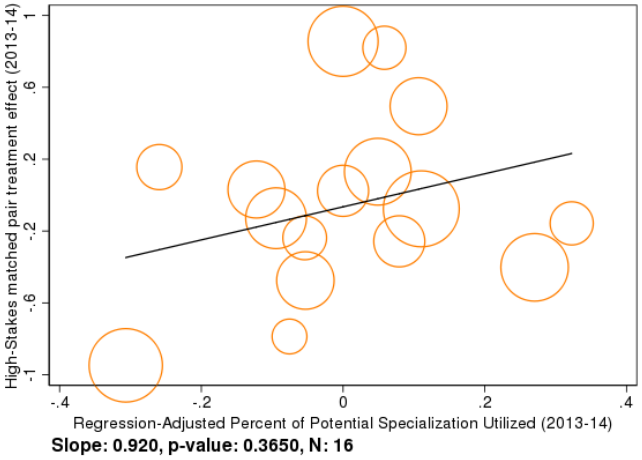


Slope: -1.348, p-value: 0.0004, N: 23

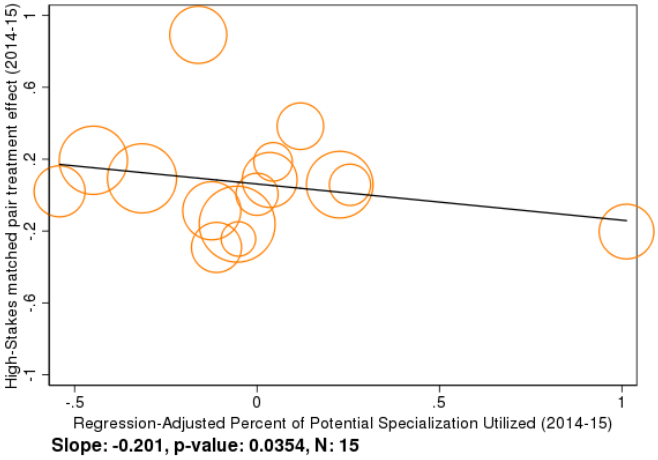


Slope: -1.246, p-value: 0.0004, N: 22

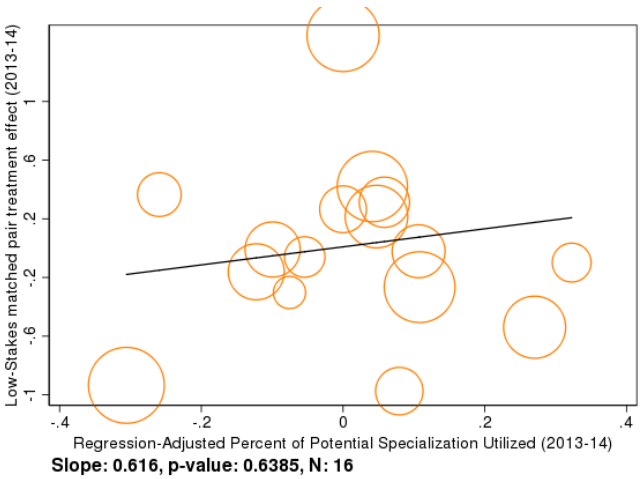
Panel C: Specialized Grade-Language Cells in Pre-Treatment Year



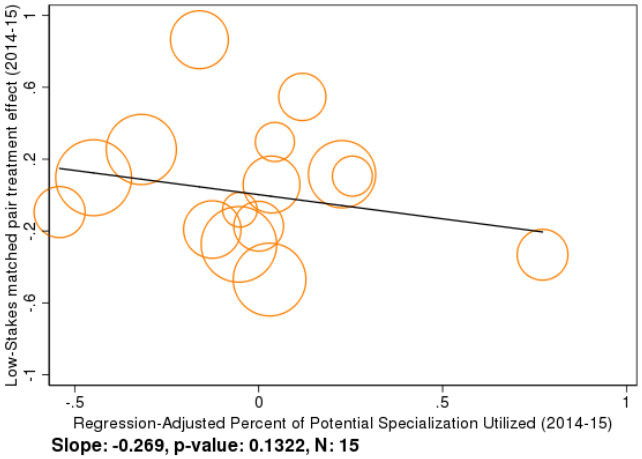
Slope: 0.920, p-value: 0.3650, N: 16



Slope: -0.201, p-value: 0.0354, N: 15



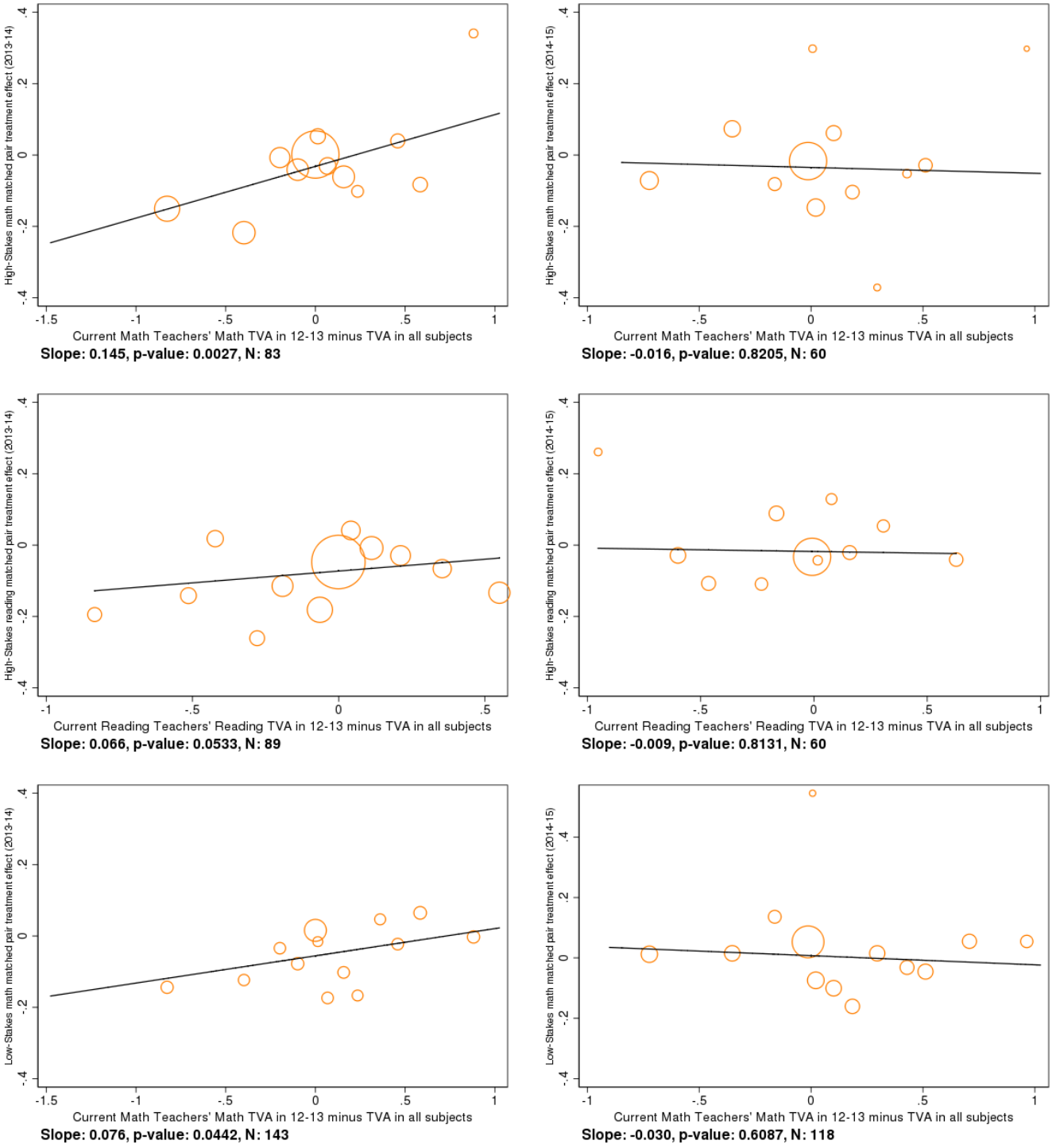
Slope: 0.616, p-value: 0.6385, N: 16

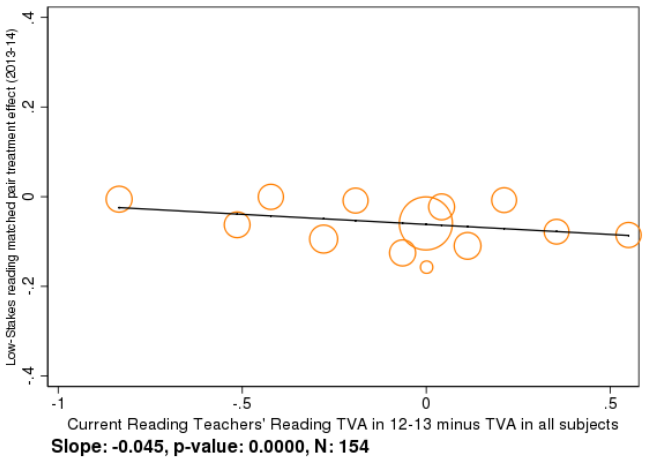


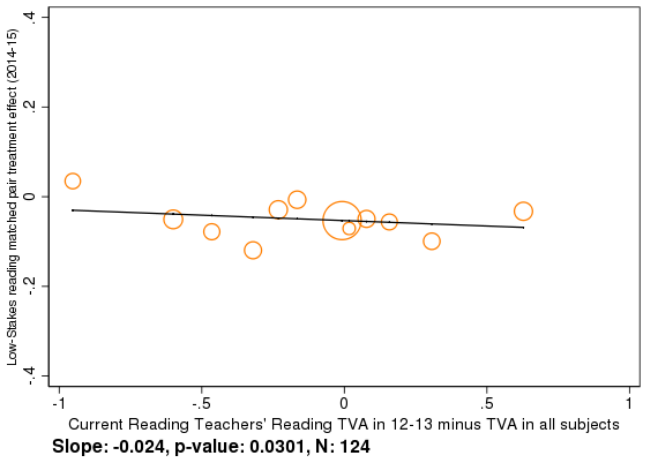
Slope: -0.269, p-value: 0.1322, N: 15

Notes: Each figure plots matched pair treatment effects against the level of specialization in a school that is induced by assignment to the treatment group. The matched pair treatment coefficient is calculated by regressing students' test scores on an indicator for treatment and controls 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 indicators for taking a Spanish baseline test as well as grade fixed effects separately for each matched pair. The level of specialization in a school is calculated via the same specification as the matched pair treatment effects with students' grade-language cells' levels of specialization as the dependent variable.

Figure 4: Teacher Average Matched-Pair Treatment Effects
and Potential Gains from Comparative Advantage







Notes: Each figure plots matched pair treatment effects and teacher characteristics. The y-axis variable is the matched pair treatment effect in either math or reading. The matched pair treatment coefficient is calculated by regressing students' test scores on an indicator for treatment and controls 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 indicators for taking a Spanish baseline test as well as grade fixed effects separately for each matched pair. The plots bin variables 15 bins, but the regression results and the line of best fit are calculated from the unbinned data.