Playing to Teachers' Strengths:
Using multiple measures of teacher effectiveness to improve teacher assignments

Lindsay Fox

Stanford University Graduate School of Education
Center for Education Policy Analysis
520 Galvez Mall
Stanford, CA 94305 USA
fox4@stanford.edu

Acknowledgements: Preparation of this manuscript by Lindsay Fox was supported in part by the Institute for Education Sciences (IES), U.S. Department of Education, through grant #R305B090016 to Stanford University. The author wishes to thank Susanna Loeb, Sean Reardon, and Chris Candelaria for their valuable feedback on earlier versions of this paper. Please direct all questions to fox4@stanford.edu.

Keywords: value-added, stability, teacher specialization, teacher effectiveness, human resource strategies

Abstract

Current uses of value-added modeling largely ignore or assume away the potential for teachers to be more effective with one type of student than another or in one subject than another. This paper explores the stability of value-added measures across different subgroups and subjects using administrative data from a large urban school district. For elementary school teachers, effectiveness measures are highly stable across subgroups, with correlations upwards of 0.9. The estimated cross-subject correlation between math and English language arts is around 0.7, suggesting some differential effectiveness by subject. To understand the magnitude of this correlation, I simulate targeted re-sorting of teachers to classrooms based on their comparative advantage. The results suggest that using multiple measures of value-added to specialize teachers by subject could produce small average increases in student achievement, and larger increases for some students.

1. Introduction

Research consistently shows teacher quality to be one of the most, if not the most, important measured within-school factors in determining student achievement (Rivkin, Hanushek, and Kain 2005). Not only do teachers meaningfully affect scholastic outcomes as well as long-term job market outcomes such as earnings (Chetty, Friedman, and Rockoff 2012), but their effectiveness varies widely as well (e.g., Aaronson, Barrow, and Sander 2007; Nye, Konstantopoulos, and Hedges 2004). Accordingly, much research is devoted to understanding and describing variation in teacher performance; however, few studies have examined whether teachers have different effects with different students or subjects. It is not difficult to imagine that one teacher may be more effective at raising achievement for one type of student than another, and, in fact, such differences have been empirically documented (e.g., Dee 2004; Loeb, Soland, and Fox 2014). Similarly, elementary school teachers, who typically teach all subjects, may be more effective in one subject than another. If teacher effectiveness does indeed vary by student type or subject, schools could raise student achievement by using multiple measures of effectiveness to match students to teachers who are the best suited to teach them.

Using eight years of rich, administrative data from a large, urban district, I estimate value-added models that allow teacher effectiveness to vary by student type or subject. I then simulate different sorting procedures, using the estimated teacher value-added to quantify the achievement gains that accrue from such sorting. In doing so, this paper examines three research questions:

1. Does teacher effectiveness, as measured by value-added, vary within a teacher by subject, specifically math and reading?
2. To what extent would re-assigning teachers to classrooms based on their value-added scores in math and reading result in increased student achievement?

3. Does teacher effectiveness, as measured by value-added, vary within a teacher by student gender, ability, ethnicity, free lunch status, or race?

The paper proceeds as follows: Section 2 motivates the paper and reviews the literature on differential teacher effectiveness, and Section 3 describes the data that is used in this study. Section 4 then details the empirical strategy for estimating teacher effects, explains the re-assignment simulations, and hypothesizes about likely results. Section 5 presents the study's findings, and Section 6 concludes by discussing policy implications of the findings.

2. Background & Motivation

Numerous states and districts have adopted policies that utilize teacher quality measures as part of teacher evaluation systems (for a summary, see National Council on Teacher Quality 2011). Most of these programs use the measures to inform high-stakes personnel decisions (such as pay and tenure), and these types of policies seem to be on the rise as evidenced by initiatives such as Race to the Top and the Teacher Incentive Fund. Education researchers have also focused on teacher quality, studying how to measure teacher impacts, how to improve teacher effectiveness, and how to hire and retain effective teachers (Kane and Staiger 2012; Rockoff et al. 2011; Loeb, Kalogrides, and Beteille 2012; Hill 2007). These policies and research approaches generally assume, however, that a teacher who is effective in one domain is effective in another. There are a number of reasons we would want to explore whether and to what extent teachers are differentially effective across different students or subgroups. For one, correlating two measures of effectiveness will give some sense of whether the measures are valid, generalizable, and suited for use in teacher evaluation. If teachers are differentially effective across subgroups, for example, this implies that two teachers could have different performance ratings simply because they were assigned different types of students. Though the ratings would

be a reflection of how the teachers performed with the students they were assigned, they may not provide fair comparisons of teacher effectiveness. Two, given the use of teacher effectiveness measures in policies that use performance incentives, correlations between the measures can shed light on whether such policies are rewarding teachers who are effective across the students and subjects they teach. Lastly, looking at multiple dimensions of teacher effectiveness may help identify which teachers are best equipped to serve particular student groups or teach particular subjects so that more productive human resource decisions could be made. For example, if teachers are differentially effective across subjects, one strategy schools could employ for improving student achievement would be to assign elementary teachers to subjects based on their relative effectiveness. Such subject specialization is already occurring in some elementary schools across the nation and has risen from 5.7% of elementary grade teachers in 1999-2000 saying their classroom organization is departmentalized to 9.9% in 2007-08 (U.S. Department of Education). Playing to teachers' strengths through targeted matching is a way schools could increase student achievement that, unlike the other levers, utilizes the current labor force and does not rely on assumptions about teacher replacements, the quality and relevance of professional development, or longer-term quality improvements in the teacher labor force.

One common measure of teacher effectiveness comes from value-added models that attempt to isolate the teacher's effect from all other factors influencing achievement such as student background characteristics, peer effects, and school characteristics. Several studies offer evidence that value-added measures from certain models may be confounded and that attaching causal interpretations to these estimates can be problematic. Some of these studies argue that the dynamic and non-random assignment of students to teachers leads to inconsistent and biased estimation of value-added measures (Rothstein 2010; Clotfelter, Ladd, and Vigdor 2006; Todd

and Wolpin 2003). Conversely, other studies contend that value-added estimates are very close to what is found from experimental assignment and that internal validity assumptions are not violated too severely (Kane and Staiger 2008; Chetty, Friedman, and Rockoff 2012). On the whole, there is no consensus on the best measure of a teacher's effectiveness, but value-added has benefits over other measures. First, unlike value-added, most conventional uses of observational protocols do not adjust for context (e.g., Pianta, La Paro, and Hamre 2008). Next, value-added directly measures student learning, reduces many forms of bias (Rubin, Stuart, and Zanutto 2004), and addresses student-teacher sorting more directly than most other measures of teacher effectiveness (Rothstein 2009; McCaffrey 2012). Lastly, with the increased availability of student test scores, value-added scores are relatively inexpensive to calculate for some teachers.

Though value-added estimates have become a popular measure of teacher effectiveness, as mentioned above, a relatively untested assumption underlies its use: that teacher effects are homogenous across student subgroups (Reardon and Raudenbush 2009). Similarly, we do not know much about whether teacher effects are homogenous across subjects. Both questions have been sparsely investigated despite their policy relevance and empirical evidence suggesting that teachers may be differentially effective.

At the elementary level where most teachers are responsible for multiple subjects, it may be the case that some teachers are stronger in teaching one subject versus another, perhaps due to college training, special certifications, or subject-specific professional development. Surprisingly, however, little empirical research has addressed this directly. In a working paper, Koedel and Betts (2007) look at the correlation of teacher value-added estimates between math and reading using elementary school panel data. The authors estimate the lower bound of the

correlation to be .35 and the upper bound to be .64. New York City data was also used to estimate the correlations across subjects and these correlations ranged from .4 to .55, though these correlations are not adjusted for sampling error (Value-Added Research Center 2010). Similarly, a paper by Loeb, Kalogrides, and Beteille (2012) report unadjusted cross-subject correlations in their appendix of between .21 and .58 depending on the value-added model used. Finally, the papers that most closely resemble the work being done in this paper are by Goldhaber, Cowen, and Walch (2013) and Condie, Lefgren, and Sims (2014). Both papers use panel data from North Carolina elementary schools to estimate value-added in math and reading and find that after correcting for sampling error, the correlations within years is around .7-.8. Taken together, these findings indicate that on average, teachers who are good at teaching math are good at teaching reading, and vice versa, but there are some differences.

In terms of relative teacher effectiveness across student groups, there is some evidence supporting the idea that some teachers may be differentially effective with one group of students or another. To start, two papers by Dee (2004, 2007) explore the effect of congruence of student-teacher race and gender on student achievement. In his 2004 paper, Dee finds that student achievement is positively affected for students who are paired with a same-race teacher. Similarly, Dee (2007) shows that assignment to a same-gender teacher significantly improves achievement as well as teacher perceptions and student engagement. Another study, by Hanushek et al. (2005), stratifies students by prior score and finds modest correlations (between .3 and .6) of average gains made by students who had the same teacher. Furthermore, a study of English Learners in New York City looks within a teacher's classroom and finds that teachers with EL-specific training and native language fluency are better with English Learners than others (Master et al. 2012). In addition to the studies cited above, a few studies look at the idea of

differential effectiveness in a value-added framework. Four studies address heterogeneity of teacher effects by student subgroups: Aaronson, Barrow, and Sanders (2007), Lockwood and McCaffrey (2009), Condie, Lefgren, and Sims (2014), and Loeb, Soland, and Fox (2014). In the first study, the authors construct separate value-added measures for teachers with a variety of different subgroups and generally find no evidence of differential effectiveness. In the second study, McCaffrey and Lockwood build a model of student-teacher interactions that allows them to test whether teacher effects are homogenous across student ability. While they find that teachers are differentially effective with students of different ability, the effects are small, explaining less than 10 percent of the overall teacher effect. Condie et al. create separate value-added measures for teachers with high-ability and low-ability students, as defined by being in the top third and bottom third of their prior school-grade test score distribution. They find correlations across ability types of .97 for math and between 0.8-0.9 for reading. Finally, Loeb et al. create separate value-added measures for teachers with English Learners and non-English Learners and find that, while teachers who are good with one group tend to be good with the other group, there are some differences. In summary, the literature generally supports the notion that teachers may be more or less effective with certain subgroups, though those differences are likely small.

In light of the relative sparseness of the literature around relative teacher effectiveness, the contributions of this paper are three-fold. One, the paper documents the extent to which teacher effects vary across subject or student type in a different context than has been studied in the literature. In particular, the paper uses a rich dataset from Miami-Dade County Public Schools, one of the largest school districts in the country, to provide another point of reference for prior findings that include correlations for multiple student subgroups as well as cross-subject

correlations. Estimating the cross-subject correlations is a particularly nice contrast to the Goldhaber et al. (2013) and Condie et al. (2014) papers for a number of reasons. The demographic composition of students is much different in Miami than in North Carolina; for example, a much larger proportion of elementary students in Miami come from low-income households than North Carolina (69 percent are eligible for free/reduced price lunch versus roughly 45 percent, respectively). The standardized tests that are administered and used for value-added calculations in both places are also different. Knowing whether the results hold across such different populations and different tests contributes to the generalizability of the findings.

The second contribution of the paper is its exploration of the strategy of targeted matching of students and teachers. Only two of the studies in the literature (Goldhaber, Cowan, and Walch 2013; Condie, Lefgren, and Sims 2014) considers a simulation of the type carried out in this study. Generally, the exercise assigns teachers to subjects based on the difference between their value-added scores in math and reading and estimates the expected achievement gains. In addition to such simulations, my paper examines additional re-assignment approaches and shows that re-assigning only the most relatively effective teachers achieves the majority of the gains from full re-sorting but with much less disruption.

Last, the paper provides a mathematical derivation of the expected gains when teachers specialize in a specific subject. No other paper on teacher relative effectiveness has addressed this question. If schools know how highly correlated value-added scores between math and reading are for their teachers, they can use the derivation to get a general sense of what the potential gains could be from subject specialization. The paper also acknowledges that principals do not have perfect information with which to sort teachers and provides another derivation

which incorporates the reliability of the value-added estimates to scale down the potential achievement gains.

3. Data

The data for this study come from the Miami-Dade County Public Schools (M-DCPS) district from the 2004-05 through 2011-12 school years for 4th and 5th grade. M-DCPS is the fourth largest school district in the nation with 392 schools and 345,000 students. The district's size as well as its relatively low teacher turnover rate makes it well suited to employing value-added analyses.

Several datasets are combined to compile the final analytic data file. First, I use an administrative database that provides student demographic data such as race, gender, free or reduced-price lunch eligibility, special education status, whether the students are limited English proficient, and the number of absences and suspensions. Second, I combine the student demographic data with test score data. In M-DCPS, students take the Florida Comprehensive Assessment Test (FCAT) and these scores comprise the entirety of the test data used in my analyses. I focus only on math and reading scores for this paper because those tests are given to all students in grades 3-5, ensuring that I have the necessary scores for the current year as well as a prior year score. Within each grade-year-subject combination, I standardize all scores to have a mean of zero and a standard deviation of one. Lastly, I use a database of courses taught by each teacher and course enrollment records to link individual students to teachers.[1] Each classroom has a unique identifier which also allows me to generate classroom-level measures such as average test scores from the prior year and percent of students eligible for free or reduced price

[1] M-DCPS uses these linked data for multiple purposes including class-size verification, student attendance, and employee evaluation. The links are kept carefully, regularly verified by teachers, and considered satisfactory for high-stakes purposes.

lunch. A unique school identifier similarly allows me to construct school-level measures. I generate classroom- and school-level characteristics using all available student-level variables, and these aggregate controls are all included in the value-added models. The only student-level variable that I construct from the data is whether a student is low ability or not. Students with prior test scores below 0 are coded as low ability, while those with a score of 0 or above are not. Panel 1 of Table 1 shows student-level means and standard deviations of the characteristics of the students in my analytic sample. A majority of the students are Hispanic and a majority are eligible for free- or reduced-price lunch. Panel 2 shows the characteristics of the teachers in the sample. The racial breakdown of teachers is mixed, with forty percent being Hispanic. Additionally, almost forty percent of teachers have an advanced degree and the average number of years of experience is 13.

4. Methods

Value-Added model

To answer my research questions, I employ a value-added model to empirically test for heterogeneous effects within teachers across subjects and student types. There are many ways of estimating value-added though no consensus exists on which is best, and addressing all of the limitations of value-added models is outside the scope of this paper; here, the focus is on whether teacher effects are invariant across subjects or student types.

I estimate value-added using a three-level hierarchical linear model. The random coefficients model is especially useful given the nested structure of the data and that it allows teacher effects to vary by subject or subgroup, which is necessary to address my research

questions.[2] The main benefit to using a random effects model instead of one of the fixed effects models is that the covariance of the teacher effects, my main parameter of interest, is estimated directly from the model, thus allowing me to perform hypothesis testing. In addition, the maximum likelihood procedure takes sampling error into account, and thus the estimate of the covariance parameter is an estimate of the true covariance of the teacher effects. For the individual teacher effects, I utilize post-estimation empirical Bayes shrinkage, which takes into account the fact that some teachers have more "information" with which to estimate their value-added simply because they had more students of a given type in their classroom. Those teachers whose effects are estimated less precisely are "shrunk" towards the grand mean across all teachers in proportion to the unreliability of the teacher-specific effect. Lastly, the random coefficients model is more efficient than fixed effects, because it does not estimate individual teacher effects, so it uses far fewer degrees of freedom. The random coefficients model assumes that the error is not correlated with the other variables in the model and that the distribution of teacher effects is bivariate normal.

In the multi-level model, level 1 is student observations, level 2 is teachers, and level 3 is schools. The primary feature of my model is that I allow teachers to have different effects for different subjects. To do so, I first construct a variable called $TestCombine$ that takes values of a student's standardized math score when the subject is coded as math, and the student's standardized reading score when the subject is coded as reading. $TestCombine$ is the dependent variable. Second, I constrain my model to have no intercept, but include both indicator variables

[2] Though I present results from the random coefficients model, I have run the models using a fixed effects specification (using the felsdvregdm command in STATA). Specifically, I estimate models with teacher and school fixed effects separately for each subgroup and then correlate the resulting value-added measures. The results are very similar and can be found in Table A.1 in Appendix A.

for math and reading, each with a coefficient that contains a random teacher effect. The random teacher effects are assumed to have a mean of zero and a variance to be estimated.

To arrive at the final model, I conducted a series of specification checks. For one, I interacted all of the controls with subject. Such interactions allow for the effects of the various controls to be different for math and reading, which is necessary since both subjects are included simultaneously in the model. An F-test with the null hypothesis of joint insignificance of the interaction terms was strongly rejected. Similarly, deviance tests support the inclusion of multiple random teacher and school effects rather than a single teacher effect and single school effect. The model for subject heterogeneity is shown below:

Level 1:

$$\begin{aligned} TestCombine_{ijst} &= \beta_{1js}(Math_{ijst}) + \beta_{2js}(Read_{ijst}) + (Math_{ijst})X_{it}'\phi_M + (Read_{ijst})X_{it}'\phi_R \\ &+ (Math_{ijst})\psi_{tM} + (Read_{ijst})\psi_{tR} + (Math_{ijst})\pi_{gM} + (Read_{ijst})\pi_{gR} \\ &+ (Math_{ijst})C_{jt}'\zeta_M + (Read_{ijst})C_{jt}'\zeta_R + (Math_{ijst})S_{st}'\omega_M \\ &+ (Read_{ijst})S_{st}'\omega_R + e_{ijst} \end{aligned}$$

where $e_{ijst} \sim N(0, \sigma^2)$

Level 2:

$\beta_{1js} = \gamma_{10s} + u_{1js}$

$\beta_{2js} = \gamma_{20s} + u_{2js}$

where $\begin{pmatrix} u_{1js} \\ u_{2js} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_1 & \tau_{2,1} \\ \tau_{1,2} & \tau_2 \end{pmatrix}\right)$

Level 3:

$\gamma_{10s} = \delta_{100} + \kappa_{10s}$

$\gamma_{20s} = \delta_{200} + \kappa_{20s}$

where $\begin{pmatrix} \kappa_{10s} \\ \kappa_{20s} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{10} & \tau_{20,10} \\ \tau_{10,20} & \tau_{20} \end{pmatrix}\right)$

The outcome variable, $TestCombine_{ijst}$, is the standardized test score for student i in classroom j in school s at time t in math or reading. The variables 'Math' and 'Read' are both binary variables that take on a value of 1 if the test score is from the corresponding subject and 0 if not. The vectors X_{it}, C_{jt}, and S_{st} contain student-, class-, and school-level covariates, respectively (see Appendix B for a complete list of variables included). $\psi_t$and π_g are year and grade fixed effects, respectively. Lastly, I generate within-school estimates, as these are more policy-relevant; it is unlikely that districts could re-assign teachers across all the schools in their district. I obtain the within-school estimates by centering the level-1 variables at the school level.

The main parameter of interest in my model is $\tau_{1,2}$which gives an estimate of the true covariance between a teacher's effectiveness in math and reading. When two variables are measured with error, as are the teacher random effects, their correlation will be attenuated. As discussed above, the attenuation bias due to sampling error is avoided by estimating the teacher effects simultaneously in a hierarchical linear model that employs maximum likelihood estimation.

The model also estimates τ_1 and τ_2 which are estimates of the true variance of the teacher effects for math and reading, respectively. Using post-estimation Bayes shrinkage, I also obtain fitted values of u_{1js} and u_{2js}, which represent the teacher- and subject-specific value-added measures. For example, u_{11s} would represent teacher 1's value-added for math in school s. Note that the model does not estimate effects by year; it uses all of the available data to estimate an overall effect for math and reading, for all teachers. The choice to predict overall teacher effects versus teacher effects by year is driven mainly by the desire to avoid simulation results that reflect idiosyncratic year-to-year variation in teacher effectiveness rather than true teacher effectiveness.

For the subgroup heterogeneity models, I run a similar model, except that the models are all estimated separately for math and reading. I investigate subgroups broken down by gender, race, ethnicity, ability, and free and reduced price lunch eligibility.

Testing the Correlations

To answer research questions 1 and 3, I test whether the correlation between teacher effects is equal to 1. A correlation of 1 would indicate that the same teachers who are effective with one group or subject are effective with the other group or subject, suggesting no differential effectiveness. Conversely, if we can reject the null hypothesis that the correlation is 1, that tells us that there are differences in effectiveness by subgroup or subject. Complications arise, however, when constructing confidence intervals around such a parameter. In particular, the confidence intervals are likely not symmetric because the normality assumptions needed to justify a symmetric confidence interval are unlikely to hold when the parameter is bounded such as in the case of a correlation or covariance, which are bounded between -1 and 1 (Raudenbush 2013). Normality is especially hard to justify if the estimate of the parameter lies near the boundary, as is the case in this analysis.

One solution to this complication is to use the delta method to construct asymmetric confidence intervals. This method essentially transforms the parameter so that it is no longer bounded, computes a symmetric confidence interval for the new unbounded parameter, and then applies the inverse transformation to that confidence interval to obtain the confidence interval of the original parameter. I employ the delta method for these analyses.

Simulations

To address research question 2, I run simulations that take advantage of any differential effectiveness observed in the data. In particular, I investigate the potential effect on student

achievement that would be realized if students are re-assigned to teachers based on multiple dimensions of value-added.

The simulations using subject-specific teacher effects are based on ranking teachers by their comparative advantage in math which is constructed by subtracting a teacher's value-added in reading from his or her value-added in math. Teachers are only re-sorted within year-school-grade cells and classrooms are kept intact. I then rank teachers on the comparative advantage measure, with the highest comparative advantage in math at the top of the list. Next, I order the original sections of math and reading with the math sections first, and then teachers are matched to the sections. For example, suppose there are two 4^{th} grade teachers in a school, and teacher A has a class of 18 students and teacher B has a class of 22 students. If teacher A has a higher comparative advantage in teaching math, I assign that teacher to teach math only; thus teacher A will teach his or her original 18 students in math as well as teacher B's original 22 students. Conversely, teacher B will now teach reading to both classes. It is possible that both teachers have a comparative advantage in math, but the one with the highest comparative advantage will teach math. It is also possible for some teachers to not be re-assigned. In the case of only one teacher in a year-school-grade cell, clearly there is no re-assignment possible. In a case where there is an odd number of teachers, one teacher will remain teaching both math and reading. Finally, there may be some year-school-grade cells where specialization is already occurring, so if the teacher who is better in math is teaching math and vice versa, there won't be any re-assignment.

This study explores the use of a comparative advantage sorting procedure in a number of ways. First, I re-assign teachers as described above using their unstandardized value-added estimates. Second, I use their standardized estimates. The difference in variance between the

math and reading scores causes the rankings to be different under these two sorting schemes. Third, I identify teachers who are in the top and bottom 5% of comparative advantage across the sample and then perform the re-assignment only within year-school-grade cells with those teachers. Lastly, I show the effects of specialization using a mathematical derivation of the expected gains from re-sorting. The details and hypotheses related to each of these procedures are discussed in the next section.[3]

Hypotheses

The first simulation ranks teachers based on their comparative advantage using unstandardized value-added scores. Comparing unstandardized value-added measures to maximize total achievement will preference the subject in which it is easier to achieve gains, namely math. This means that a principal will be trading off higher gains in math for lower and possibly negative gains in reading. Figure 1 shows the intuition for this graphically: because math scores have higher variance and the value-added scores in math and reading are positively correlated, better overall teachers will be assigned to math. Thus, total gains will be positive due to teachers being assigned to subjects in which they are relatively better, but reading gains will be sacrificed for higher overall gains driven by math.

The second simulation uses standardized value-added in math and reading. Specifically, the value-added are standardized to have a mean of 0 and a standard deviation of 1 within schools. Doing so puts math and reading on the same scale such that one subject is not favored over the other during the re-assignment; in other words, a one standard deviation gain of the standardized math value-added is as valuable as a one standard deviation gain of the standardized

[3] Similar simulations were planned for student subgroups but are not included because the results (discussed in the Findings section and shown in Table 6) do not indicate meaningful differences in teacher effectiveness across student subgroups.

reading value-added. While reading gains are sacrificed in the first simulation, total gains are sacrificed in the second simulation; using the standardized value-added essentially imposes an additional constraint that gains in teacher value-added be balanced across the subjects, so we expect that total gains would be smaller than in the first simulation which capitalizes on the fact that it is easier to generate gains in math.

Another way to utilize multiple measures of value-added but not impose specialization on everyone is to only re-assign the best and worst teachers based on their comparative advantage. Most of the gains from re-sorting are coming from tails of the distribution, so this strategy is a way to still realize gains while minimizing the amount of re-sorting. If one cost to specializing teachers is the disruption it causes to students, this is a potential sorting mechanism that takes that into account and places value on doing less re-sorting. This strategy can be applied when using the standardized or unstandardized value-added scores.

Lastly, the paper provides a mathematical derivation of the gains that could be expected from subject specialization. The derivation yields a comparison between random assignment of teachers to classrooms and optimal assignment of teachers to classrooms. In the derivation, all teachers who are better at math than reading are assignment to math and vice versa. The results therefore present a different comparison than the simulations which show the contrast between current and feasible-optimal sorting.

5. Findings

The first research question asks whether teachers are differentially effective across the subjects they teach. Table 2 shows the variance of teacher effects for math and reading as well as the correlation between them. Consistent with the literature, there is greater variance in teacher effects for math than for reading. The correlation between math and reading effectiveness is .715

which is statistically significantly different from 1. This finding suggests that teachers who are effective in math tend to be effective in reading and vice versa, but that there is some leverage for targeted student-teacher matching to translate into achievement gains.

To understand the magnitude of the within-teacher variation in effectiveness across subjects, I carry out a number of simulations. The first simulation re-sorts all teachers (when possible) and assigns them to classes of students based on their comparative subject advantage using unstandardized value-added measures. Table 3 shows these results. I estimate total gains of .061 standard deviations, while gains in math are .096 and gains in reading are -.035.[4] As expected, gains are higher in magnitude for math than reading, and we do indeed see negative gains in reading. Because it is not possible to re-assign all teachers, only 85% of students have a new teacher under this simulation. For just those students who are re-assigned, the gains are .072 standard deviations. Table 3 also shows the results of the simulation where schools are only re-sorted if they have a teacher from the top or bottom 5% of teachers. This simulation results in only 45% of students having a new teacher. Total gains are estimated to be .035 standard deviations, while gains are .058 in math and -.023 in reading. For only those students who are re-sorted, the total gains are .078 standard deviations. To compare the two simulations in percentage change metrics, we have reduced the number of students who are re-assigned by almost 50%, but gains were reduced by only about 40%. Thus, by only reassigning teachers who are relatively very effective at math or reading, modest gains can still be realized.

Table 4 shows the gains from simulations in which the sorting is done using standardized value-added measures. As discussed above, this strategy acknowledges that the variation in test scores in math and reading are quite different, and places more value on a balance between the

[4] The expected gains from specialization could be different if un-shrunk estimates of teacher value-added are used, though it is more likely that shrunk estimates would be used in practice.

two subjects. The total gains from using the standardized value-added scores are .016 in math and .014 in reading, for a total of .030 standard deviations. As expected, reading gains are no longer negative, but the overall magnitude of gains is lower than when un-standardized scores are used for re-assignment. To put the gains in context, a gain of .016 in math is roughly 3% of the average annual achievement gains in math (Hill et al. 2008) and a gain of .014 in reading is roughly equivalent to half of the students not having a first-year teacher (Rockoff 2004). Lastly, gains of .020 can still be achieved by only re-sorting the top and bottom 5% of teachers based on their comparative advantage, and the gains for only the students who are re-sorted in this simulation are .046.

The results in Tables 3 and 4 contrast current with feasible-optimal sorting and assume that principals have perfect information with which to assign teachers to classrooms. Table 5 provides a contrast between random assignment of teachers to classrooms and optimal sorting of teachers to classrooms, allowing for the reliability of the value-added measures to be imperfect. The numbers in Table 5 are derived according to Appendix C with parameter values set to those reported in Table 2. As the reliability of the value-added measures decreases, the value of using them for sorting is similarly diminished. For example, with perfectly reliable value-added, sorting teachers on their standardized value-added would lead to a .07 standard deviation gain in math and a .04 gain in reading; but when the reliability of the measures is set to 0.7, gains drop to .05 and .03 standard deviations, respectively. All else equal, districts that have more years of data will estimate value-added with higher reliability and can thus expect higher gains from specialization.[5]

[5] In addition to the reliability of the estimates, the gains one would expect from sorting may also be affected by real year-to-year variation in value-added due to other reasons such as gains from experience or professional development.

My findings about within teacher heterogeneity across subjects are remarkably similar to those reported in Goldhaber et al. (2013) and Condie et al. (2014) despite the demographic differences between Miami-Dade and North Carolina as well as the different assessments used in each setting. Both studies find cross-subject correlations bewteen 0.7-0.8, where my estimated correlation is around 0.7. The magnitude of the gains from specialization are also quite similar. For example, total gains using unstandardized value-added in Miami are .061 standard deviations, and they are .063-.082 in North Carolina (Goldhaber et al. 2013). The North Carolina results, however, come from simulations with schools of exactly four classrooms such that everyone can be re-sorted. Gains from Miami for those who are re-sorted are .073 which falls in the middle of the reported range in North Carolina. Small differences also emerge due to differences in the estimated variances of the math and reading value-added scores, but plugging in estimated parameters from North Carolina into my derivation yields results very close to those reported by Goldhaber et al.[6] Using standardized value-added, Condie et al. (2014) report gains of 0.054 and 0.030 standard deviations in math and reading, respectively, for a correlation of 0.7. These values are also well within the range of the derivation results reported in the second panel of Table 5. These similarities show that even in very different contexts, school systems that specialize elementary teachers have the potential to modestly increase student achievement. In both settings, however, there are other factors that might affect the magnitude of the gains. From a developmental perspective, we might worry that there could be negative consequences of structural changes that result from implementing subject specialization in the elementary grades. For example, Eccles and her colleagues argue that, among other variables, teacher-student relationships decrease during the transition to middle school where the school environment is

[6] For instance, if I plug in values reported in Goldhaber et al. (2013) for the parameters of the derivation in Appendix C, the derivation yields gains of .092 and -.004 for math and reading with reliability=1, while Goldhaber et al. report gains of .092 and -.013 for perfect information with the apparently random assignment sample.

markedly different (Eccles et al. 1993). Conversely, teachers who are now assigned to teach only one subject may be able to generate higher value-added since they can now allocate a bigger portion of their lesson planning time to one subject. Gains might also be underestimated if teachers are happier teaching only the subject in which they excel, resulting in increased teacher retention. Future research could test the validity and importance of such hypotheses.

The last question this paper asks is whether teacher effectiveness varies within a teacher by student subgroup. As was foreshadowed in the methods section, I do not find meaningful differences by subgroup. Table 6 shows the estimated correlation between the teacher effects for all of the models. Across math and reading, all of the subgroup correlations are above .9, with some as high as .99.[7] These findings suggest that there is hardly any differential effectiveness within teachers by student gender, ability, race and free lunch status. It is possible, however, that the way in which students are sorted to classrooms could bias the results. For example, if principals are sorting students to teachers based on an unobservable trait that is correlated with a subgroup characteristic and differential teacher effectiveness, this could attenuate the extent of differential effectiveness for that subgroup. On the other hand, if principals are sorting based on an unobservable trait and effectiveness with a subgroup *across* teachers, the sign of the bias is ambiguous. For example, if teacher A is better than teacher B with students with behavioral problems, and the principal sorts all of the students with behavioral problems to teacher A, the correlations for, say, low ability/high ability students could be underestimated if teacher A is equally effective with all students. Ultimately, the correlations are quite high suggesting that gains from any type of re-sorting would be negligible, so I do not explore this issue in greater detail.

[7] The reported correlations are only based on teachers who have students in both groups. Over 80% of teachers have students in both groups for female, low ability, free lunch, and Hispanic, but for white and black, the percentages are 57 and 66, respectively.

6. Conclusion

The goal of this study is to test whether teacher effectiveness is heterogeneous within teachers and to explore the importance of such heterogeneity if it exists. In particular, if teachers are differentially effective, human resource strategies could play to teachers' strengths by using multiple performance measures to make better teacher assignments. My findings suggest that there is substantial stability in teacher effectiveness across student subgroups, but that there is less stability across subjects. The correlations of teacher value-added scores across subgroups are all above 0.9, while the correlation across subjects is around 0.7. Thus, it appears that value-added to these student exams is capturing largely the same underlying teaching ability, though slightly less so when comparing the value-added to the math and reading exams. If schools were to leverage the fact that there are some teachers who are differentially effective by subject, it would be possible to increase average achievement scores on the exams through teacher specialization. The gains from such a policy heavily depend on the correlation between math and reading value-added and the reliability of the value-added estimates. This study's parameter estimates show that small gains in student achievement can be realized in a system that specializes elementary teachers by subject.

Caution should be taken when thinking about implementing such a strategy, however. First, aspects of teaching that are believed to lead to desired non-test outcomes may not be closely related to value-added measures of teacher effectiveness (Bill and Melinda Gates Foundation 2011); thus, the benefit of re-sorting on such measures might lead to increased achievement on the student assessment, but not to longer-run student outcomes such as increased college completion or higher wages. Similarly, value-added estimates for teachers can vary across different exams (Papay 2011). If states or districts change the student exam, it is unclear

whether or how much of the gains would still be accrued- a particularly relevant caveat as states continue to roll out new Common Core assessments.

Though my study looks at the effect of specialization based on value-added estimates, many other measures of teacher quality could be used and the underlying idea still holds: if teachers are relatively more effective in some domains than others, schools can use that information to make more productive human resource decisions. Unlike other policy levers, the gains from specializing teachers are achieved using the current labor force and do not rely on assumptions about longer-term labor market shifts or the efficacy of professional development. Specializing teachers simply uses information from multiple measures of effectiveness to help manage the current stock of teachers more efficiently and put them in classrooms where they are most likely to succeed with their students.

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics* 25 (1):95-135.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2012. The Long-term Impacts of Teachers: Teacher Value-added and Student Outcomes in Adulthood. NBER Working Paper No. 17699.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. Teacher-Student Matching and the Assessment of Teacher Effectiveness. NBER Working Paper No. 11936.

Condie, Scott, Lars Lefgren, and David Sims. 2014. Teacher heterogeneity, value-added and education policy. *Economics of Education Review* 40:76-92.

Dee, Thomas S. 2004. Teachers, Race, and Student Achievement in a Randomized Experiment. *The Review of Economics and Statistics* 86 (1):195-210.

Dee, Thomas S. 2007. Teachers and the Gender Gaps in Student Achievement. *The Journal of Human Resources* 42 (3): 528-554.

Eccles, Jacquelynne, Carol Midgley, Allan Wigfield, Christy Miller Buchanan, David Reuman, Constance Flanagan, and Douglas Mac Iver. 1993. Development During Adolescence: The Impact of Stage-Environment Fit on Young Adolescents' Experiences in Schools and in Families. *American Psychologist* 48 (2):90-101.

Foundation, Bill and Melinda Gates. 2011. Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project. Seattle, WA: Bill and Melinda Gates Foundation.

Goldhaber, Dan, James Cowan, and Joe Walch. 2013. Is a good elementary teacher always good? Assessing teacher performance estimates across subjects. *Economics of Education Review* 36:216-228.

Hanushek, Eric A., John F. Kain, Daniel M. O'Brien, and Steven G. Rivkin. 2005. The Market for Teacher Quality. NBER Working Paper No. 11154.

Hill, Carolyn J, Howard S Bloom, Alison Rebeck Black, and Mark W Lipsey. 2008. Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives* 2 (3):172-177.

Hill, Heather C. 2007. Learning in the Teaching Workforce. *The Future of Children* 17 (1):111-127.

Kane, Thomas J., and Douglas O. Staiger. 2008. Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. NBER Working Paper No. 14607.

Kane, Thomas J., and Douglas O. Staiger. 2012. Gathering feedback for teachers: Combining observations with student surveys and achievement gains. Policy and Practice Brief prepared for the Bill and Melinda Gates Foundation.

Koedel, Cory, and Julian R. Betts. 2007. Re-Examining the Role of Teacher Quality In the Educational Production Function. Working Paper.

Lockwood, J. R., and Daniel F. McCaffrey. 2009. Exploring Student-Teacher Interactions in Longitudinal Achievement Data. *Education Finance and Policy* 4 (4):439-467.

Loeb, Susanna, Demetra Kalogrides, and Tara Beteille. 2012. Effective Schools: Teaching Hiring, Assignment, Development, and Retention. *Education Finance and Policy* 7 (3):269-304.

Loeb, Susanna, James Soland, and Lindsay Fox. 2014. Is a Good Teacher a Good Teacher for All? Comparing Value-Added of Teachers With Their English Learners and Non-English Learners. *Educational Evaluation and Policy Analysis*. In Press.

Master, Ben, Susanna Loeb, Camille Whitney, and James Wyckoff. 2012. Different Skills: Identifying Differentially Effective Teachers of English Language Learners. National Center for Analysis of Longitudinal Data in Education Reserach.

McCaffrey, Daniel F. 2012. Do value-added methods level the playing field for teachers? In *What We Know Series: Value-Added Methods and Applications*, edited by Carnegie Foundation for the Advancement of Teaching.

National Council on Teacher Quality. 2011. State of the States: Trends and Early Lessons on Teacher Evaluation and Effectiveness Policies.

Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges. 2004. How Large Are Teacher Effects? *Educational Evaluation and Policy Analysis* 26 (3):237-257.

Papay, John P. 2011. Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates Across Outcome Measures. *American Educational Research Journal* 48 (1):163-193.

Pianta, Robert C, Karen M La Paro, and Bridget K Hamre. 2008. Classroom assessment scoring system. *Baltimore: Paul H. Brookes*.

Raudenbush, Stephen W. 2013. Delta Method for HLM. Personal Communication.

Reardon, Sean F, and Stephen W Raudenbush. 2009. Assumptions of value-added models for estimating school effects. *Education Finance and Policy* 4 (4):492-519.

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

Rockoff, Jonah E. 2004. The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review* 94 (2):247-252.

Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger. 2011. Can You Recognize an Effective Teacher When You Recruit One? *Education Finance and Policy* 6 (1):43-74.

Rothstein, Jesse. 2009. Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy* 4 (4):537-571.

Rothstein, Jesse. 2010. Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics* 125 (1):175-214.

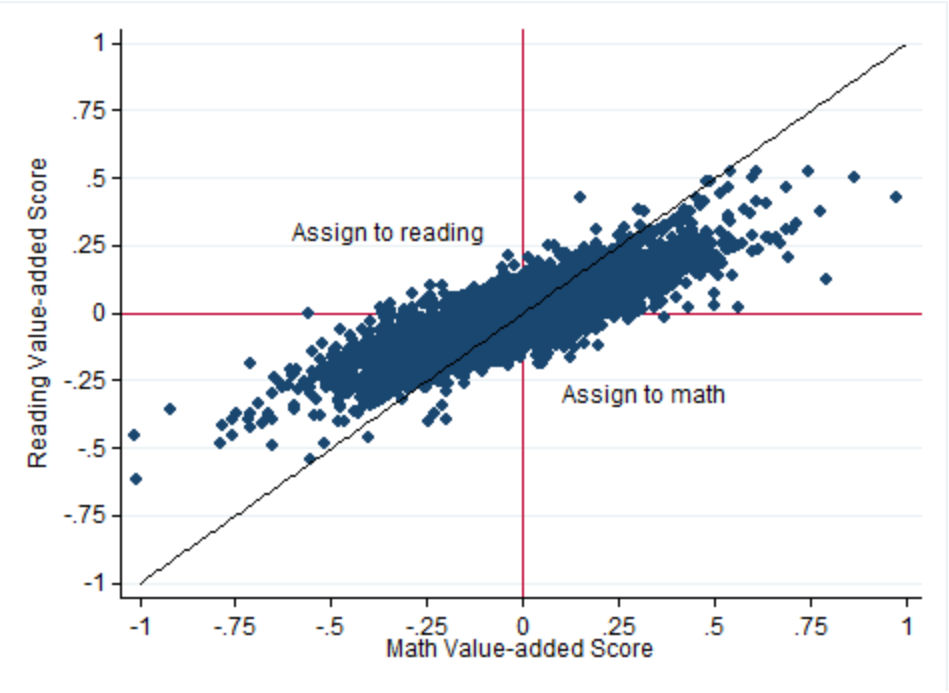
Rubin, Donald B., Elizabeth A. Stuart, and Elaine L. Zanutto. 2004. A Potential Outcomes View of Value-Added Assessment in Education. *Journal of Educational and Behavioral Statistics* 29 (1):103-116.

Todd, Petra E., and Kenneth I. Wolpin. 2003. On the Specification and Estimation of the Production Function for Cognitive Achievement. *The Economic Journal* 113:F3-F33.

Value-Added Research Center. 2010. NYC Teacher Data Initiative: Technical Report on the NYC Value-Added Model. University of Wisconsin-Madison.

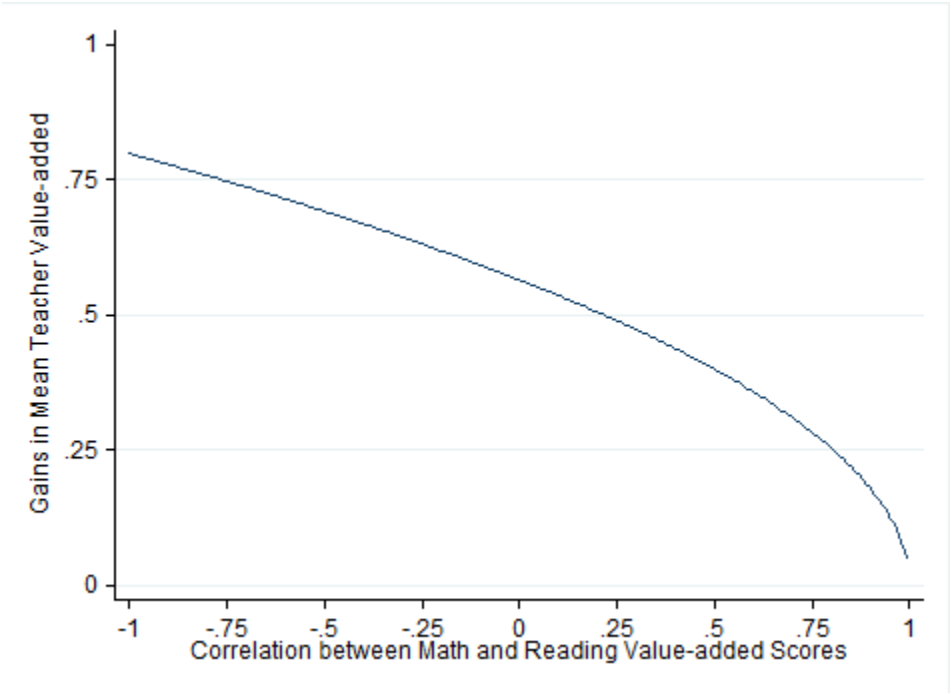
Figures

Figure 1. Subject Value-added Assignment Rule



Note. Dots represent individual teachers and the black line represents where math value-added and reading value-added are equal. In this stylized version of re-assignment, all teachers below the line are assigned to math and all those above the line are assigned to reading.

Figure 2. Relationship between Cross-subject Correlation and Gains in Teacher Value-added



Note. The curve shows the relationship between the correlation of math and reading value-added scores and gains in mean standardized teacher value-added when all teachers who are better in math than reading are assigned to math, and vice versa. For the full derivation, see Appendix C.

Tables

Table 1. Summary Statistics for Analytic Sample

	Mean	Standard Deviation
Student-Level Descriptives		
Math Score	0.081	(0.944)
Reading Score	0.093	(0.928)
White	0.095	(0.293)
Black	0.246	(0.431)
Hispanic	0.633	(0.482)
Other	0.025	(0.157)
Female	0.494	(0.500)
Free/Reduced Price Lunch	0.685	(0.464)
Limited English Proficient	0.082	(0.274)
Special Education	0.110	(0.313)
Retained	0.004	(0.067)
Absences in prior year	6.117	(6.161)
Student Suspensions in prior year	0.086	(0.781)
Number of student-year observations	347,913	
Teacher-Level Descriptives		
White	0.301	(0.459)
Black	0.277	(0.448)
Hispanic	0.409	(0.492)
Other	0.013	(0.112)
Female	0.890	(0.313)
Advanced Degree	0.394	(0.489)
Experience	13.142	(9.202)
Number of teacher-year observations	17,318	

Note. Test scores are standardized to be mean zero with unit standard deviation by year, grade, and subject before students are matched to teachers.

Table 2. Standard Deviation and Correlation of Math and Reading Teacher Value-added Estimates

	Estimate (95% CI)
Math standard deviation	0.2244
Reading standard deviation	0.1368
Correlation between math and reading	0.7151 (.6965,.7327)

Note. Standard deviations and correlation of the value-added are estimated directly from a random effects model. 95 percent confidence intervals are estimated using the delta method.

Table 3. Achievement Gains from Subject Specialization using un-standardized value-added measures

	Re-sorting All Teachers Average Gain	Re-sorting top and bottom 5% Average Gain
Math	0.0958	0.0580
Reading	-0.0349	-0.0226
Total	0.0608	0.0354
Total for those re-sorted	0.0725	0.0778

Note. Total gains are the average of the sum of students' math and reading test scores.

Table 4. Achievement Gains from Subject Specialization using standardized value-added measures

	Re-sorting All Teachers Average Gain	Re-sorting top and bottom 5% Average Gain
Math	0.0161	0.0116
Reading	0.0135	0.0084
Total	0.0296	0.0200
Total for those re-sorted	0.0351	0.0458

Note. Total gains are the average of the sum of students' math and reading test scores.

Table 5. Derived Achievement Gains from Subject Specialization

Unstandardized Value-added (math gains/reading gains)

		Reading reliability			
		1	0.9	0.8	0.7
Math reliability	1	0.145 -0.015	0.139 -0.015	0.132 -0.014	0.125 -0.013
	0.9	0.131 -0.014	0.126 -0.013	0.122 -0.013	0.116 -0.012
	0.8	0.118 -0.013	0.115 -0.012	0.111 -0.012	0.107 -0.011
	0.7	0.106 -0.013	0.104 -0.011	0.101 -0.011	0.098 -0.010

Standardized Value-added (math gains/reading gains)

		Reading reliability			
		1	0.9	0.8	0.7
Math reliability	1	0.069 0.043	0.063 0.039	0.058 0.036	0.052 0.032
	0.9	0.063 0.039	0.059 0.036	0.054 0.034	0.050 0.031
	0.8	0.058 0.036	0.054 0.034	0.051 0.031	0.047 0.029
	0.7	0.052 0.032	0.050 0.031	0.047 0.029	0.050 0.031

Note. Expected effects of specialization are derived according to Appendix C with parameters from Table 2.

Table 6. Standard Deviation and Correlation of Teacher Value-added by Student Type

	Math estimate (95% CI)	Reading estimate (95% CI)
Female standard deviation	0.2311	0.1329
Male standard deviation	0.2297	0.1507
Correlation between female and male	0.9937	0.9886
	(.9931,.9943)	(.9873,.9898)
Low Ability standard deviation	0.2567	0.1644
High ability standard deviation	0.2063	0.1168
Correlation between low ability and high ability	0.9438	0.9134
	(.9383,.9489)	(.9032,.9225)
Free lunch standard deviation	0.2333	0.1483
Not free lunch standard deviation	0.2231	0.1264
Correlation between free lunch and no free lunch	0.9925	0.9802
	(.9918,.9932)	(.9778,.9824)
White standard deviation	0.2207	0.1326
Non-white standard deviation	0.2317	0.1425
Correlation between white and non-white	0.9800	0.9989
	(.9773,.9824)	(.9986,.9991)
Black standard deviation	0.2351	0.1558
Non-black standard deviation	0.2298	0.1368
Correlation between black and non-black	0.959	0.9284
	(.9546,.9630)	(.9186,.9372)
Hispanic standard deviation	0.2325	0.1385
Non-Hispanic standard deviation	0.2286	0.1468
Correlation between Hispanic and non-Hispanic	0.9696	0.9488
	(.9667,.9722)	(.9433,.9538)

Note. Standard deviations and correlations of the value-added are estimated directly from a random effects model. 95 percent confidence intervals are estimated using the delta method.

Appendix A:

Table A.1. Teacher value-added correlations using a fixed effects specification

	Math estimate	Reading estimate
By student type		
Correlation between female and male	0.9918	1.0001
Correlation between low ability and high ability	0.9391	0.9440
Correlation between free lunch and no free lunch	0.9526	0.9383
Correlation between white and non-white	0.9518	0.9509
Correlation between black and non-black	0.9356	0.9334
Correlation between Hispanic and non-Hispanic	0.9700	0.9865
By subject		
Correlation between Math and Reading	0.6740	

Note. Correlations are disattenuated by dividing the correlation by the square root of the product of the reliabilities of each subgroup/subject. Controls are listed in Appendix B.

Appendix B: List of covariates included in models

- Prior math test score
- Prior reading test score
- Black
- Hispanic
- Other race
- Female
- Free lunch eligible
- English learner status
- Special education status
- Prior absences
- Prior suspensions
- Class average prior math test score
- Class average prior reading test score
- Class percentage black
- Class percentage Hispanic
- Class percentage other race
- Class percentage female
- Class percentage free lunch eligible
- Class percentage English learners
- Class percentage special education
- Class average prior absences
- Class average prior suspensions
- School percentage black
- School percentage Hispanic
- School percentage other race
- School percentage female
- School percentage free lunch eligible
- School percentage English learners
- School percentage special education
- School average prior absences
- School average prior suspensions
- School enrollment

Appendix C: Derivation of Gains from Subject Specialization

From the random coefficients model, the following parameters are estimated: $\tau_M, \tau_{MR}, \tau_R$.

$$\begin{pmatrix} u_{jM} \\ u_{jR} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_M & \tau_{RM} \\ \tau_{MR} & \tau_R \end{pmatrix}\right)$$

We want to know what happens if every teacher who has a u_M greater than u_R is assigned to teach only math, and vice versa with reading. To figure out gains in math, we want to know $E[u_M | u_M > u_R]$, so we can integrate over the joint probability density function:

$$\frac{\int_{-\infty}^{+\infty} \int_{u_R}^{+\infty} u_M f(u_M, u_R) du_M du_R}{\int_{-\infty}^{+\infty} \int_{u_R}^{+\infty} f(u_M, u_R) du_M du_R}$$

where $f(u_M, u_R) = \frac{1}{2\pi\sigma_M\sigma_R\sqrt{1-\rho^2}} e^{\frac{-\left(\frac{{u_M}^2}{\sigma_M^2} + \frac{2\rho u_M u_R}{\sigma_M\sigma_R} + \frac{y^2}{\sigma_R^2}\right)}{2(1-\rho^2)}}$

By transforming some of our variables, we can turn the double integral into a single integral from which we can derive a closed form solution.

Let $s = u_M - u_R$ and $t = u_M$. In matrix notation, $A = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}$ and $\begin{pmatrix} s \\ t \end{pmatrix} = A \begin{pmatrix} u_M \\ u_R \end{pmatrix}$. Then we know that $\begin{pmatrix} s \\ t \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, A\tau A^T\right)$ and we want to know $E[t | s > 0]$.

Define $\tilde{t} = t - cs$, where $c = \frac{E[ts]}{E[s^2]} = \frac{Cov(t,s)}{Var(s)}$ and $s \sim N(0, \sigma^2)$.

This transformation makes it so that $\tilde{t}$ and s are independent, so we can simplify the integral from a double integral to a single integral. So now we want to know $E[\tilde{t} + cs | s > 0]$.

$E[\tilde{t} + cs | s > 0] = E[\tilde{t}] + E[cs | s > 0]$ since $\tilde{t}$ and s are independent.

$E[\tilde{t}] + E[cs | s > 0] = E[cs | s > 0] = cE[s | s > 0]$ since we know $E[\tilde{t}] = 0$.

$$cE[s | s > 0] = \frac{c \int_0^{+\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} dx}{\int_0^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} dx}$$

$$= 2c \int_0^{+\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} dx$$

$$= -c\sigma \sqrt{\frac{2}{\pi}} e^{\frac{-x^2}{2\sigma^2}} \Bigg|_0^{+\infty}$$

$$= c\sigma \sqrt{\frac{2}{\pi}}$$

$$= \sqrt{\frac{2}{\pi}} \frac{\tau_M - \tau_{MR}}{\sqrt{\tau_M + \tau_R - 2\tau_{MR}}}$$

$$= \sqrt{\frac{2}{\pi}} \frac{\tau_M - \rho_{MR}\sqrt{\tau_R/\tau_M}}{\sqrt{\tau_M + \tau_R - 2\rho_{MR}\sqrt{\tau_R/\tau_M}}}$$

where ρ_{MR} is the correlation between u_M and u_R. If $\rho_{MR} \geq 0$, then:

a) if $\tau_R \leq \tau_M$, then $E[u_M | u_M > u_R] > 0$.

b) if $\tau_M < \tau_{MR}$, then $E[u_M | u_M > u_R] < 0$.

If the variances aren't equal, then the expected value of the one with the smaller variance can be negative. In particular, it will be negative if variance of the smaller is less than the covariance.

Suppose we set $\tau_R = \tau_M$, then the gain from reassigning teachers to teach the subject they rank highest at will be a function of ρ_{MR}:

$$E[u_M | u_M > u_R] = E[u_R | u_R > u_M] = \sqrt{\frac{1 - \rho_{MR}}{\pi}}$$

If teacher effectiveness in math and reading are very highly correlated, the gains from optimal reassignment are small. However, if correlations are modest or negative, the gains can be very large.

Above, we wanted to know $E[u_M|u_M > u_R]$. Due to measurement error in value-added, however, it may be more relevant to ask about $E[u_M|\hat{u}_M > \hat{u}_R]$. We can proceed in a similar fashion as above, with a few adjustments.

First, let $s = \hat{u}_M - \hat{u}_R = u_m - u_r + e_m - e_r = u_m - \tilde{u}_r$. Now, $\tilde{\tau} = \begin{pmatrix} \tau_m & \tau_{mr} \\ \tau_r & \tau_r + \omega_m^2 + \omega_r^2 \end{pmatrix}$, assuming that $cov(u, e) = 0 \; and \; cov(e_m, e_r) = 0$. As before, let $t = u_M$; then, $A = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}$ and $\begin{pmatrix} s \\ t \end{pmatrix} = A \begin{pmatrix} u_m \\ \tilde{u}_r \end{pmatrix}$. Then we know that $\begin{pmatrix} s \\ t \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, A\tilde{\tau}A^T\right)$ and we want to know $E[t|s > 0]$. All of the steps proceed as above, and the result is:

$$E[u_M|\hat{u}_M > \hat{u}_R] = \sqrt{\frac{2}{\pi}} \frac{\tau_M - \rho_{MR}\sqrt{\tau_R/\tau_M}}{\sqrt{\tau_M + \omega_m^2 + \tau_R + \omega_r^2 - 2\rho_{MR}\sqrt{\tau_R/\tau_M}}}$$

The reliability of the math and reading value-added scores is the respective signal to noise ratios, defined as: $r_m = \frac{\tau_m}{\tau_m + \omega_m^2}$ and $r_r = \frac{\tau_r}{\tau_r + \omega_r^2}$, respectively. We can then re-write the formula as:

$$= \sqrt{\frac{2}{\pi}} \frac{\tau_M - \rho_{MR}\sqrt{\tau_R/\tau_M}}{\sqrt{\tau_m/r_m + \tau_r/r_R - 2\rho_{MR}\sqrt{\tau_R/\tau_M}}}$$

When math and reading value-added scores are standardized, such that $\tau_R = \tau_M = 1$, then the result is:

$$= \sqrt{\frac{2}{\pi}} \frac{1 - \rho_{MR}}{\sqrt{1/r_m + 1/r_R - 2\rho_{MR}}}$$