Assessing Teacher Quality for Disadvantaged Students*

Joshua Hollinger[†] Alexis Orellana[‡]

November 8, 2021

Abstract

Does lower teacher quality contribute to short-term and long-term outcomes for disadvantaged students? We leverage transfers of elementary teachers across schools in North Carolina to measure differences in teachers' effects on contemporaneous and future test scores according to students' socio-economic characteristics. We quantify the importance of these differences to account for the observed test score gaps between disadvantaged and advantaged students. Variation in teacher quality accounts for 3% of the total variation in contemporaneous test scores. We also find that teacher quality accounts for similar proportions when we consider variability in test scores taken two and three years after. Our estimates are robust to bias-correction methods that account for limited mobility bias.

Keywords: inequality, teacher quality, achievement gaps

*We thank John Singleton for providing us access to the data stored in the North Carolina Education Research Data Center. We also thank Lisa Kahn, Ronni Pavan, and Kegon Tan for helpful comments and suggestions.

[†]Department of Economics, University of Rochester. E-mail: jhollin3@ur.rochester.edu

[‡]Department of Economics, University of Rochester. E-mail: borellan@ur.rochester.edu

1 Introduction

Differences in education quality across ethnic or racial groups have been a salient topic in the United States due to their implications in segregation, intergenerational mobility patterns, and, ultimately, unequal distribution of opportunities. Despite substantial efforts to reduce the gaps in access to equal educational opportunities, differences are still quite large. Reardon et al. (2019) document that the dispersion in racial/ethnic achievement gaps across school districts is huge in the U.S., ranging from nearly zero to 1.5 test score standard deviations. Since research has shown that the allocation of teachers across schools is one of the most relevant inputs in the educational process, addressing existing gaps in teacher quality has been of interest to governments and policymakers.

This paper revisits the differences in teacher quality received by different groups, employing data from elementary students in North Carolina public schools between 1997 and 2011 to estimate teacher value-added on short- and long-term outcomes. We contribute to the previous literature on this topic by addressing two new elements. First, while much of the literature has focused on differences in teacher quality based on contemporaneous value-added or observable characteristics (Clotfelter et al., 2005; Sass et al., 2012; Goldhaber et al., 2015; Mansfield, 2015), recent research has emphasized the importance of teachers for longer-run outcomes (Jackson, 2018; Petek and Pope, 2021; Gilraine and Pope, 2020). We incorporate this dimension into our analysis by assessing to what extent disadvantaged students are also exposed to teachers with lower capabilities to increase test scores in subsequent years. Following the literature, we refer to this capability as *long-term* value-added. Understanding these differences is relevant from a policy perspective. First, to the extent that under-served students are also exposed to teachers with lower capabilities to increase learning in the long run, estimates of differences in contemporaneous value-added will underestimate the total impact of the current teacher allocation. Second, ignoring this dimension can induce an incomplete or inaccurate evaluation of teachers. Under time and resource constraints, teachers decide between short-term and long-term learning practices. A strong focus on end-of-year achievement tests can induce teachers to prefer the former (e.g., rote memorization, teaching to the test) while ignoring other longer-term strategies (e.g., improve study habits, teach foundational concepts). To the best of our knowledge, no previous work has attempted to analyze disparities in teacher quality alongside this additional temporal dimension.

As a second contribution, we study differences in teacher effectiveness at the elementary level. The closest work to ours is Mansfield (2015), who analyzes how teacher quality varies at the high-school level in North Carolina. He finds that variation in teacher quality accounts for a small fraction of the overall variability in ninth-grade test scores.[1] In this paper, we focus on fourth-grade and fifth-grade students in North Carolina, for whom we observe math and English end-of-grade test scores until eighth grade. By following these students until middle school, we can quantify the long-run implications of teacher allocation in elementary grades. Since teachers can increase skills

[1] Specifically, he uses end-of-course tests for English 1, Econ/Law/Politics, US History, Algebra 1, Algebra 2, Geometry, Biology, Chemistry, Physics, and Physical Science.

relevant for outcomes observed in middle and high school, an emphasis on allocating more effective teachers to relatively more disadvantaged students at this level could be more productive to reduce performance gaps observed in later years.

The rest of the paper proceeds as follows: Section 2 describes the data that we use in this study. Section 3 details our empirical framework for estimating the importance of teacher quality. Section 4 discusses how we adapt recent contributions in the employee-employer literature to our context. Section 5 presents our main findings. Finally, Section 6 concludes.

2 Data

We use data on all fourth-grade and fifth-grade students in North Carolina public schools between 1997 and 2011. The data, provided by the North Carolina Education Research Data Center (NCERDC), contains information about state-level test scores and a large number of student and teacher characteristics. Student characteristics include past test scores in third and fourth grades, gender, race, free/reduced lunch status, parental education, and behavioral measures (number of absences, repetition status). Teacher characteristics include gender, race, experience, and education. We focus on students taking fourth or fifth grade for the first time.

We observe teacher identifiers linking teachers and students in the end-of-grade datasets. Nevertheless, the person administering the test is not necessarily the teacher who taught that class. To construct reliable links between teachers and classrooms, we employ additional information in the personnel data, including school, subject, grade, class size, and aggregate demographic information (race-gender totals) for each class taught by a teacher. We implement a fuzzy matching algorithm used in previous work (Jackson, 2014; Mansfield, 2015; Jackson, 2018) to match classrooms observed in the end-of-course and the personnel datasets. The procedure is the following: we compute the number of students by gender, race, and grade for each classroom in the end-of-course and the personnel datasets. Classes that match across files based on school, subject, grade, teacher identifier, class size, and demographic information are considered perfect matches. For the remaining classrooms, we compute a weighted distance between classrooms using these variables. Then, we select the pairs of classrooms which satisfy our minimum-distance criterion. Following this procedure, we match 80% and 85% of the total number of student observations for English and math, respectively.

2.1 Connected Schools

To separately identify the effects of schools and teachers, we require that teachers appear in more than one school over time. We leverage teacher transfers to obtain this variation and get a connected graph of schools, where schools correspond to nodes and teachers correspond to edges linking schools. For our purposes, we define a transferring teacher as one who has taught at least twenty students in two different schools in different years. In addition, we drop teachers who taught less than fifty students in a given school to allow more precise estimates of the teacher fixed effects. Starting with Abowd et al. (1999) (hereafter AKM), the empirical literature in labor markets using longitudinal employer-employee data has applied the same intuition to separately identify firm from worker unobserved characteristics.[2] Papers using this type of models in education are Jackson (2013), Mansfield (2015), and Thiemann (2019). To get a higher number of connected schools, we pool information for fourth-grade and fifth-grade students taking math or English classes between 1997 and 2011. In our main analysis, we employ our set of two-edge connected sets for each subject independently. We use the common set of schools to obtain pooled estimates. We restrict our set of connected schools to the subset satisfying the following two conditions: First, we require at least five transferring teachers between each pair. Second, the subset is two-edge connected, meaning that each pair of schools connects by at least two combinations of edges that do not have any edge in common. After imposing these restrictions, we obtain a set of 471 and 474 two-way connected schools for English and math courses, respectively. After pooling observations for both subjects, we get 498 schools.

Table 1 shows some statistics for our final sample, considering a set of 498 two-edge connected schools. Our overall sample consists of around 800,000 students with valid math and reading scores in either fourth or fifth grade and 8,700 teachers. 50% of all students are female, 57% are white, 31% are black, 6% are Hispanic, and 2% are Asian. 44% of students are classified as free lunch or reduced-price lunch. About 39% of the students in our sample belong to a family where the highest level of education is high school or below, and 26% where the highest educational level is at least a four-year college degree. The average class size in our sample is twenty students.

3 Empirical Framework

Let y_{ijsgt} be an outcome observed for a fourth-year or fifth-year student i taught by teacher j in grade g and school s during year t. We focus on elementary students since our data includes end-of-grade contemporaneous and middle school test scores. Specifically, we consider standardized test scores observed for each student in year t, as well as in years $t+2$ and $t+3$. We assume that the production function depends on observable student and classroom characteristics, X_{ijgst}, and unobserved school-level inputs ϕ_s. Importantly, following the value-added literature, X_{ijgst} includes measures of the student's test scores and behaviors in prior years. Teachers contribute to student learning through their unobserved time-invariant teacher quality, μ_j, and as a function of their total years of experience, $f(\exp_j)$, and school-specific years of experience, $f^S(\exp_{js})$.

$$y_{ijsgt} = \underbrace{X'_{ijgst}\beta}_{\substack{\text{Student}\\ \text{Background}}} + \underbrace{\phi_s}_{\substack{\text{School}\\ \text{Quality}}} + \underbrace{\mu_j + f(\exp_{jt}) + f^S(\exp_{jst})}_{\text{Teacher Quality}} + \underbrace{\epsilon_{ijsct}}_{\substack{\text{Unexplained}\\ \text{Variation}}} \tag{3.1}$$

[2] See, for example, Card et al. (2013) using German data, and Card et al. (2015) using Portuguese data

We estimate this specification pooling the information for each student in English and math subjects. To allow for grade heterogeneity in our pooled estimates, we interact each element of X_{ijgst}, $\exp_{jt}$, and $\exp_{jst}$ with a fifth-grade indicator. We also estimate (3.1) separately by subject. X_{ijgst} includes an indicator for free or reduced-price lunch status, indicators for parent's education, indicators for each combination of race and gender, third-degree polynomials for lagged test scores, the interaction of math and English lagged scores. Since we observe end-of-grade tests from third grade, we observe two lagged scores only for fifth-grade students. We use the end-of-grade scores in third grade and tests taken at the start of that grade for fourth-grade students. We also include the number of absent days and an indicator for suspensions in the previous grade. Since a student's learning can also be affected by the characteristics of their classmates, we include leave-one-out class averages of all of these variables as well, along with class size. We allow $f(\exp_{jt})$ to be a flexible function of total experience. We employ indicators for zero, one to two, three to five, six to ten, eleven to fifteen, sixteen to twenty, twenty to thirty, and more than thirty years of experience. To proxy for school-teacher match effects, we incorporate an additional function $f^S(\exp_{jst})$ which denotes the number of years a teacher has served in the same school. We use indicators for zero, one to two, three to five, six to ten, and more than ten years.

Specification (3.1) closely resembles Mansfield (2015), who studies differences in teacher quality for high-school students in North Carolina. Nevertheless, our specification differs from his since he leverages school-course transfers to identify school and teacher fixed effects separately. Instead, we employ only school fixed effects, so our empirical strategy considers teachers observed teaching the same subject at different schools. We discuss these differences in more detail in the next subsection. This value-added specification allows us to quantify the importance of each input to explain the overall variation of each outcome y_{ijsgt}. We employ this framework to analyze the importance of teacher inputs in producing short- and long-term outcomes and differences in the quality of these inputs across different sub-groups of students. Our focus on elementary schools allows us also to analyze differences in the production of skills that have long-term relevance for middle school outcomes.

An important focus of this paper is to analyze the variance in teacher inputs that persist after a student has been taught to teacher j. In this sense, when y_{ijsct} denotes a future outcome, our estimates of $\hat{\mu}_j, \hat{f}$ and $\hat{f}^S$ will capture a combination of short-term and long-term components of value-added (Jacob et al., 2010; Gilraine and Pope, 2020). Short-term value added refers to effects in test scores derived from transitory learning or teaching to the test. On the other hand, long-term value added, can be associated with persistent gains in learning if a teacher, for example, motivates students to study more effectively or teaches skills required in later grades.[3]

[3]To make this point clear, assume that the teacher fixed effect can be decomposed as $\mu_j = \delta^S \mu_j^S + \delta^L \mu_j^L$ where μ_j^S and μ_j^L correspond to short-term and long-term value-added, respectively. Gilraine and Pope (2020) assume $\delta^S = 0$ to estimate the teacher component in year $t + 1$

3.1 Identification

In a model with teacher and school fixed effects, identification requires teacher transfers and schools to form a connected graph, where teachers connect school cells. In other words, any two schools in the set must be connected following a path of teachers transferring across schools. Our approach differs from Mansfield (2015) in that we also analyze math and English separately and use school fixed effects rather than school-course fixed effects. Therefore, we identify teacher quality solely based on teachers teaching the same subject in different schools, rather than using the variation based on teachers switching subjects.

To consider the assumptions required to rule out endogeneity concerns that would violate identification, we assume that the error component ϵ_{ijsgt} can be decomposed into three components:

$$\epsilon_{ijsgt} = \theta_{st} + \zeta_{jt} + \upsilon_{ijsgt} \tag{3.2}$$

The first term, θ_{st}, represents factors that vary by school and year, such as time-varying school quality, a change in school disciplinary policy, a shock affecting a school's test scores in a given year, or changes in the unobserved quality of school's student body over time. The second term, ζ_{jt}, represents factors that vary by teacher and year, such as improvement in an individual's teaching ability that is unrelated to experience, a shock affecting a teacher's performance in a specific year, or fluctuations in performance due to contexts which are suited better or worse to a teacher's skills, beyond the heterogeneity allowed by $f(\exp_{jt})$. The third term represents factors specific to the individual student's skill measure, which are not captured by X_{ijgst}.

The fundamental identification assumption is that teacher mobility is conditionally random, that is, unrelated to any component in ϵ_{ijsgt}. The first potential endogeneity problem involves a correlation between teacher mobility and θ_{st}. This correlation may arise if teachers choose to move to schools for which θ_{st} is increasing, or if they choose to leave schools for which θ_{st} is decreasing. Examples include good teachers choosing to transfer when a bad principal comes into their school, transfers driven by a sudden influx of resources into a school, and transfers driven by changes in unobserved student characteristics within a school.

The second potential endogeneity problem involves a correlation between teacher mobility and ζ_{jt}. If a teacher's performance appears better or worse due to moving to a better or worse school, that is, changes to ϕ_s, this is captured by our school fixed effects and our estimates are properly identified. But if a teacher's performance gets systematically better or worse after transferring for any other reason, identification is violated. This could come from a better teacher-school match, a transferring teacher being assigned unobservably worse students in their first year in their new school, or teachers being more likely to transfer when they are improving faster.

4 Methodology

Based on our estimates from (3.1), we define teacher quality as:

$$\hat{\tau}_{jt} = \hat{\mu}_j + \hat{f}(\text{exp}_{jt}) + \hat{f}(\text{exp}_{jst}) \tag{4.1}$$

We start by showing the results of variance decompositions that rely on our restriction to two-edge connected schools and the conditionally exogenous mobility assumption to identify school and teacher effects separately. One concern when estimating two-way fixed effects models is that limited mobility of teachers across schools can bias our estimates of the variance of teacher quality. This issue resembles the problem observed in AKM models using employee-employer data, where the presence of a limited number of workers connecting firms can generate biased estimates of the relative contribution of workers and firms to wage variability, as well as biased estimates of worker-firm complementarities Bonhomme et al. (2020). In our context, if schools are weakly connected because only a few teachers move across them, estimates of school effects will be biased upwards and estimates of the sorting between schools and teachers will be biased downwards. To address this issue in particular, we apply the methodology proposed by Kline et al. (2020) to our setting to obtain bias-corrected estimates of the variance decomposition of contemporaneous and future test scores.

Kline et al. (2020) develop a bias-corrected estimator of variance components by using leave-out estimates of the error variances. This procedure involves computing a "leave-one-out" connected set of schools, corresponding to those schools that remain connected after removing any single transferring teacher. That is, given the largest connected component of a bipartite network $\mathcal{G}$ where schools and teachers are vertices, the algorithm removes all teachers who are articulation points in this graph.[4]

An alternative method to obtain bias-corrected estimates is proposed by Bonhomme et al. (2020). They use a correlated random-effects model to correct for limited mobility bias. This approach reduces the number of parameters to estimate by restricting the means and covariances of worker and firm effects. Although their model is more parsimonious and offers the advantage of being computationally more tractable, it comes at the cost of imposing restrictions that may be strong in our context. In particular, their model assumes that the covariance between firms j and j' is zero whenever $j \neq j'$ and that the cross-worker covariance of idiosyncratic shocks is zero. This restriction precludes correlation of school effects or common shocks for a group of students in the same classroom in our context. For these reasons, we employ the methodology of Kline et al. (2020), which, although computationally more demanding, is less restrictive about the covariance structure of school and teacher fixed effects.

[4]For more details, see the Data and Computational Appendix in Kline et al. (2020).

5 Results

5.1 Baseline Variance Decompositions

We start by analyzing how the overall variance of each outcome is explained by each of the components of (3.1). We refer to school quality as the estimated fixed effect ($\hat{\phi}_s$), and teacher quality as the sum of each teacher's persistent productivity plus total experience and school-specific experience, $\hat{\tau}_{jt} = \hat{\mu}_j + \hat{f}(\exp_{jt}) + \hat{f}(\exp_{jst})$, as defined in (4.1).

Table 4 shows our decomposition of the variance in contemporaneous test scores and future test scores. We pool fourth-grade and fifth-grade students in English and math classes. Column (1) presents our variance decomposition for contemporaneous test scores. This column shows that around 69% (=0.657/0.958) of the total variation is explained by underlying student and peers characteristics. In contrast, teacher quality alone explains only 2.4%, while the sum of teacher and school inputs explain 3.4%. These numbers are small and consistent with the estimates of Mansfield (2015) for high-school students. He finds that school and teacher inputs combined explain around 5% of the total variation in contemporaneous test scores. On the other hand, while he estimates that background characteristics account for 60% of the total variation, we find that this share increases to 69% for elementary students. Our estimates of the variance of teacher quality imply that moving one student from the median teacher to another one at the 84th percentile in this distribution associates to an increase of 0.15 test score standard deviations, which is in line with the estimates reported in the literature (Jackson et al., 2014).[5] We find a covariance of -0.007 between our estimates of school quality and teacher quality, $cov(\phi_s, \tau_{jt})$.

Column (2) shows the variance decomposition of test scores taken two years after being exposed to teacher j. We employ sixth-grade and seventh-grade standardized test scores in the respective subject as our outcome of interest. As expected, we find that the explained variation decreases relative to column (1). Our explanatory variables in (3.1) account for 71% of the overall variation. We also find some differences between the importance of each factor to explain contemporaneous and future achievement. Background characteristics now explain 60% (=0.568/0.953). Interestingly, teacher quality at t explains a similar fraction of the total variation in test scores in $t+2$ relative to contemporaneous test scores. $Var(\tau_{jt})$ accounts for 2.2% of the total variation of two-years-ahead test scores, while school and teacher inputs combined account for 4%.

Column (3) shows a similar pattern. This column presents the decomposition of the total variance of test scores taken three years after (at the end of seventh-grade or eighth-grade). We observe that classroom and student background characteristics explain 58% (=0.546/0.941) of the total variation, while teacher inputs account for 2.4%. Tables 5 and 6 show our estimates when we repeat this exercise separately by subject. As the previous literature reports, math teachers explain a higher fraction of test variability than English teachers.

[5] INCLUDE NUMBERS FOR HS AS WELL

We contextualize our findings using the estimates of Chetty et al. (2011) who measure by how much kindergarten classroom quality explains the variability in test at the end of kindergarten and eighth-grade. Since they observe teachers only once, their estimates of class effects combine the effects of teacher, peers, and other class-level shocks. They find large contemporaneous effects of classrooms but a quick fade-out over the following years. Their results show that an increase of 1 s.d. in kindergarten class quality increases students' contemporaneous test scores by 0.32 s.d. However, the longer-term impact on eighth-grade test scores is zero. We find that the explained variation due to teacher effects, although small, remains constant over time while the importance of student-level (lagged test scores and demographics) and classroom-level inputs decrease substantially. Therefore, our estimates suggest that this fade-out is explained mainly by the decreasing importance of lagged test scores.

The negative and small correlation between school and teacher effects, although small, suggests no sorting of teachers to schools. Nevertheless, the existing evidence suggests a non-negative sorting between schools and teachers. Lankford et al. (2002) and Hanushek et al. (2004) show that high-poverty schools with high ethnic minority enrollment shares tend to have teachers with lower qualifications than low-poverty schools. Using the end of busing policies in 2002 in North Carolina, Jackson (2009) shows that schools experiencing an increase of black students share saw a decrease in the proportion of experienced teachers, a reduction in the proportion of teachers with higher scores on licensure exams, and a decrease in teacher value-added. Therefore, one concern is that this estimate could resemble the negative correlation between firm and worker effects found in the earnings inequality literature and thus may be indicative of limited mobility bias (Bonhomme et al., 2020). Although we limit our set of schools to those connected by several transferring teachers, it could still be possible that mobility is insufficient to estimate school and teacher fixed effects appropriately. We employ two robustness checks to test whether our set of schools is sufficiently well-connected and the validity of our baseline estimates. Specifically, in section 5.4 we apply the connectivity tests of Jochmans and Weidner (2019) and the bias-correction method developed by Kline et al. (2020), which puts additional restrictions to our set of schools.

5.2 Differences in Teacher Quality by Student Background

While the estimates of Table 4 show variation across all students, we are also interested in assessing whether the allocation of effective teachers, both in terms of short-term and long-term value-added, correlates with student characteristics. To explore whether this is the case, we employ our estimates of background characteristics $X'_{ijgt}\hat{\beta}$ from (3.1) to create an index of student disadvantage. We group students by deciles and compute the average teacher quality for each one of them.[6] Figure 1 presents our results for contemporaneous, two-years-ahead, and three-years-ahead test scores using the pooled sample. Consistent with our previous discussion, the y-axis shows a small variation in the average teacher quality serving students with different backgrounds. Sub-figure (a) shows

[6] We also consider an index based only on gender, race, previous test scores, parental education, and free- or reduced-price lunch status. We find very similar results to what we present here.

that the difference between the average quality for the top and bottom deciles is 0.025, meaning that teacher quality accounts only for around 1% of the test score gap between these two groups. Considering the distribution, we observe relatively small differences for the first five deciles and much of the variation at the top of the distribution. The difference in average teacher quality between the tenth and ninth deciles is higher than between the seventh and first deciles. Subfigures (b) and (c) show similar patterns when we plot differences in long-term teacher quality for test scores taken by students at $t+2$ and $t+3$, respectively. First, we observe that the gap between the top and bottom deciles shrinks. The differences in average teacher quality between the top and bottom deciles are 0.02 and 0.015 approximately. Second, we also observe an increasing slope as we reach the top deciles of the disadvantage index distribution. We conclude that, although differences in teacher quality do not explain a considerable fraction of the total variability in test scores, we observe significant differences in its allocation by students' characteristics. Moreover, we find similar patterns for short-term and long-term teacher value-added.

Figures 2 and 3 show the relationship between teacher quality and student background separately by subject. We obtain each plot after estimating (3.1) employing only English or math teachers, and then computing a background index for each case.[7] As shown in Tables 5 and 6 we observe more variation for Math teachers. Figure 2 shows that the variation between the top and bottom decile for contemporaneous test scores is around 0.3 standard deviations. We also observe that while there is a positive gradient for subsequent periods, differences for English teachers are relevant only for the same period. Variation in teacher quality in periods $t+2$ and $t+3$ is practically zero, suggesting that differences in teacher quality have a more lasting impact on Math than for English courses.

5.3 Between-School and Within-School Differences

We conclude our analysis by decomposing the total variation in between-school and within-school differences. To this extent, we compute within-school and between-school differences in teacher quality across subpopulations. We focus on the difference between the top and bottom deciles and the top and bottom quartiles. Table 7 displays the differences in school-average teacher quality $\bar{\hat{\tau}}_s$ as well as within-school deviations $\hat{\tau}_{jt} - \bar{\hat{\tau}}_s$ for contemporaneous and future test scores. Panel A shows that the difference in contemporaneous test scores between the top and bottom deciles is 2.8 standard deviations. The difference in total teacher quality τ_{jt} (corresponding to the sum of the teacher-specific fixed effect and experience effects) between these subgroups is 0.024 standard deviations. School-level differences explain a larger fraction of this variation. While, on average, students at the top decile attend schools where the average teacher quality is 0.005 higher than those attended by students at the bottom decile, we find that this difference increases to 0.019 standard deviations when we compute within-school differences. Overall, this difference accounts for less than 1% of the total 2.8 standard deviation gap between top and bottom deciles.

[7] We compute the two-edge connected set of schools using the same restrictions applied to the pooled sample. The only difference is that we consider transferring teachers teaching the specific subject in different schools.

The comparison of differences in long-term value-added reveals a similar pattern. Panels B and C show similar magnitudes when computing the between-school and within-school differences in teacher quality when we consider middle school test scores.

5.4 Bias-Corrected Variance Decompositions

Finally, we turn to examining the robustness of our results by applying the methodology of Kline et al. (2020). We first obtain leave-one-out connected sets and then compute bias-corrected estimates of the variance of teacher quality, school effects, and their covariance. Table 8 summarizes these estimates. For our pooled sample, the set of schools used for the estimation reduces to 480. Compared to our baseline estimates of Table 4, we obtain slightly smaller estimates for the variance of teacher quality. The fraction of variation in contemporaneous test scores explained by teacher quality corresponds to 2%. The covariance between school and teacher effects reduces by half although it is still negative.[8]

6 Conclusion

In this paper, we study the distribution of teacher quality for elementary students in North Carolina and its relevance to explain test score gaps in current and future grades. Our work adds to a vast literature studying the short- and long-run consequences of teacher allocation. Our work also relates to the literature studying the persistence of teacher effects (Jacob et al., 2010; Gilraine and Pope, 2020).

We find that teacher quality gaps account for a small fraction of the differences in contemporaneous test scores. Considering math and English teachers together, the gap between students in the top and bottom deciles of our disadvantage index accounts for around 2% of the total variation in test scores. Math teachers account for a higher share. We also find that variation in contemporaneous teacher quality accounts for a similar share of the variance in test scores taken by the same group of students two and three years later when they attend middle school. Our analysis by subject shows that the impact of math teachers is the most relevant. Overall, these results suggest that the allocation of elementary teachers in North Carolina benefits more advantaged students in short- and long-term learning.

Our estimates of the covariance between teacher and school inputs show a negligible degree of sorting, even after conducting a robustness check to validate our results using leave-one-out connected sets (Kline et al., 2020). Potential reasons to explain this behavior are the limited incentive schemes present in North Carolina school districts or unobserved determinants of teacher location decisions. Nevertheless, our estimates are consistent with previous studies employing administrative data from

[8] Using the same bias-correction method, Thiemann (2019) also finds a negative but small covariance between school and teacher effects.

North Carolina (Mansfield, 2015; Thiemann, 2019). We are not aware of other studies that have quantified the importance of school-teacher sorting outside this state.

As policymakers become interested in developing more sophisticated ways to assess teachers, for example, measuring the ability to increase long-term knowledge, it is crucial to learn which practices make a difference between effective and less-effective teachers. Identifying these strategies and targeting teachers adequately is an important step towards reducing the inequality of educational inputs.

Moreover, it is not obvious how to incorporate these additional dimensions into an evaluation scheme. Teachers can allocate effort towards certain activities encouraged by the scheme, leaving others unattended. Investigating how teachers may respond to these changes and their impact on student achievement are important questions for future research.

7 Figures and Tables

Table 1: Summary Statistics.

Variable	Mean	Std. Dev.	Number Obs.
Unit of observation: Student-year			
Female	0.50	0.50	807830
White	0.57	0.50	807830
Black	0.31	0.46	807830
Hispanic	0.06	0.25	807830
Asian	0.02	0.14	807830
Other race	0.04	0.18	807830
Reading Score (lagged)	0.11	0.97	807830
Math Score (lagged)	0.12	0.98	807830
Free or Reduced Lunch	0.44	0.50	807830
Parental educ: At most high school	0.39	0.49	807830
Parental educ: At least four-year degree	0.26	0.44	807830
Class size	20.49	1.23	807830
Fourth grade	0.49	0.50	807830
Unit of observation: Teacher-year			
Total experience: 0 years	0.08	0.27	8731
Total experience: 1-5 years	0.38	0.48	8731
Total experience: 6-15 years	0.29	0.45	8731
Total experience: 15-25 years	0.15	0.36	8731
Total experience: >25 years	0.10	0.30	8731
School experience: 0 years	0.23	0.42	8731
School experience: 1-5 years	0.66	0.47	8731
School experience: >5 years	0.12	0.32	8731

Table 2: Correlation of Teacher Fixed-Effect Estimates: Math teachers

	Test Score t	Test Score $t+2$	Test Score $t+3$
Test Score t	1		
Test Score $t+2$	0.431	1	
Test Score $t+3$	0.352	0.704	1

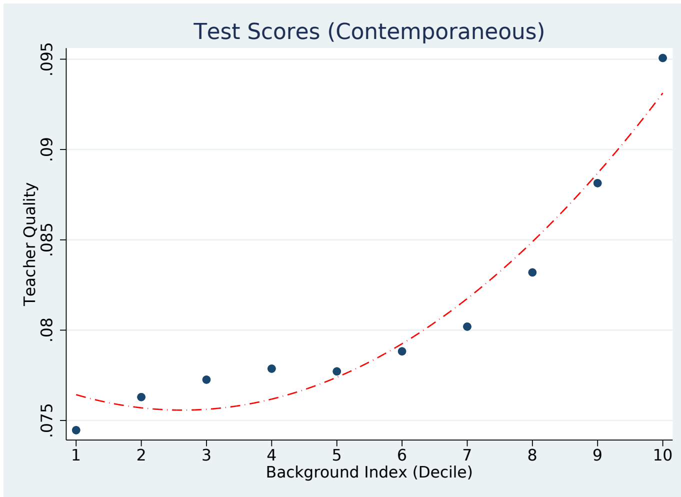
Notes: This matrix reports the correlation between the teacher fixed effects $\hat{\mu}_j$ estimated from (3.1) after using each of the outcomes as the dependent variable. $N = 8,542$ teachers.

Table 3: Correlation of Teacher Fixed-Effect Estimates: English teachers

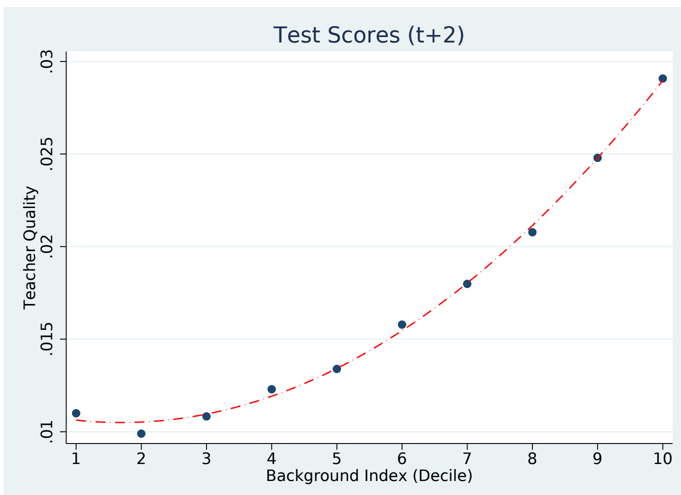
	Test Score t	Test Score $t+2$	Test Score $t+3$
Test Score t	1		
Test Score $t+2$	0.392	1	
Test Score $t+3$	0.371	0.653	1

Notes: This matrix reports the correlation between the teacher fixed effects $\hat{\mu}_j$ estimated from (3.1) after using each of the outcomes as the dependent variable. $N = 8,486$ teachers.

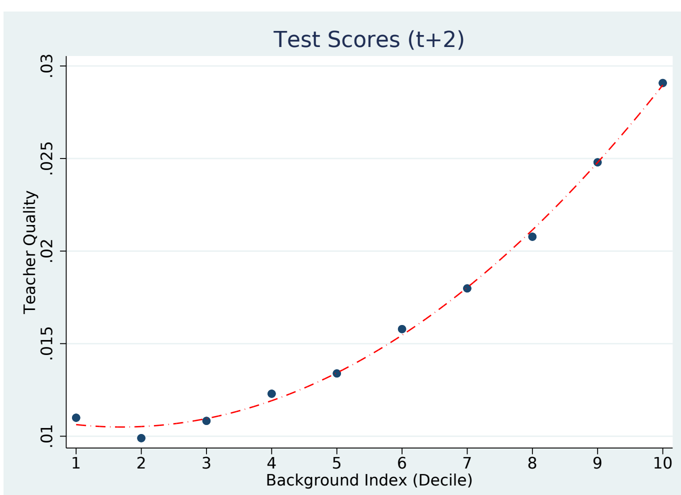
Figure 1: Differences in Teacher Quality by Background Index: Pooled Subjects



(a) Contemporaneous test scores

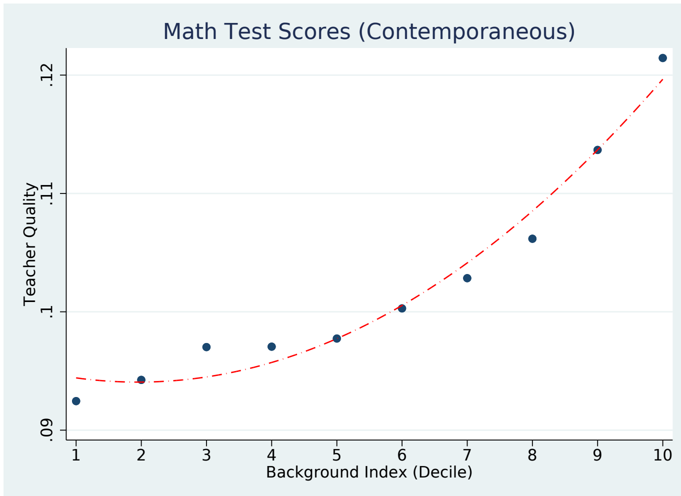


(b) Test scores at $t + 2$

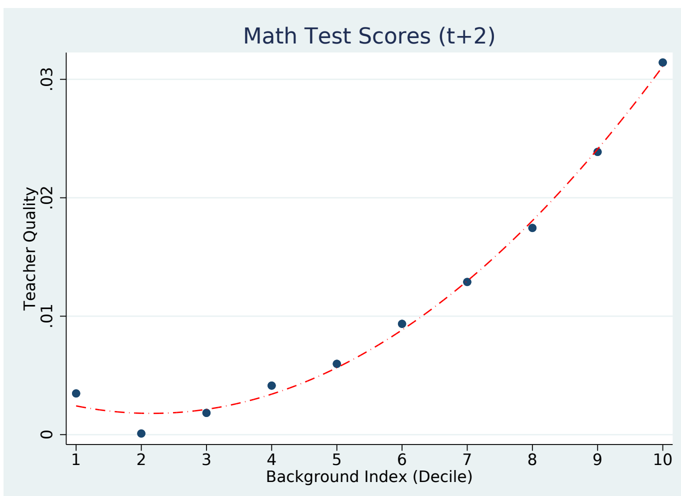


(c) Test scores at $t + 3$

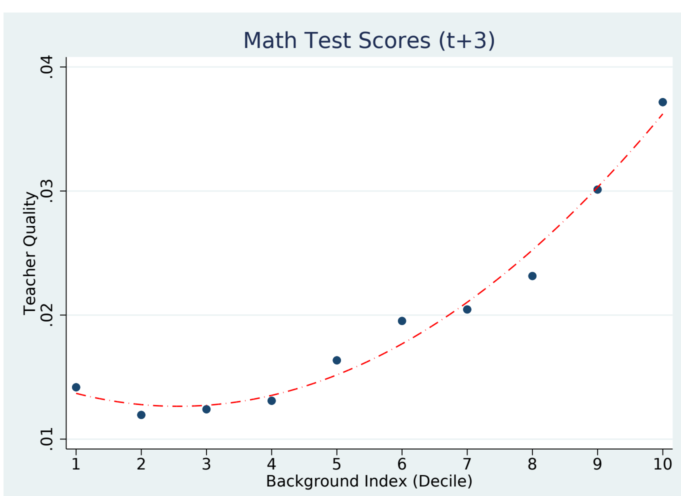
Figure 2: Differences in Teacher Quality by Background Index: Math



(a) Contemporaneous test scores

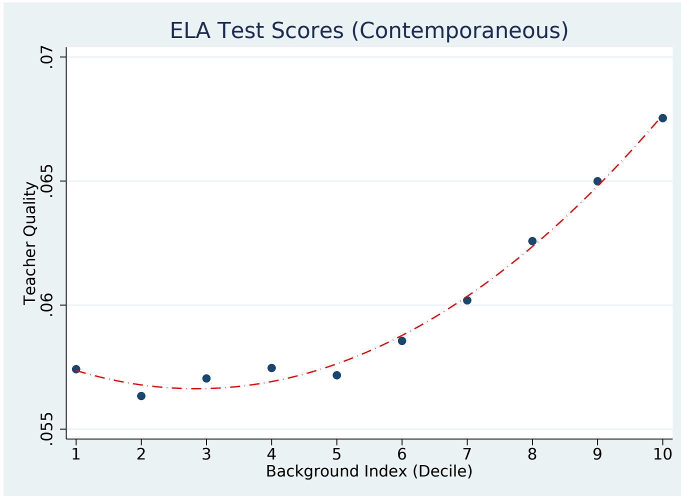


(b) Test scores at $t+2$

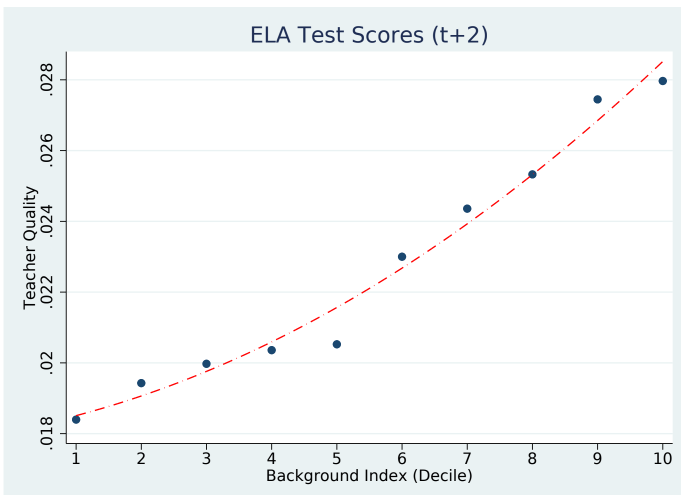


(c) Test scores at $t+3$

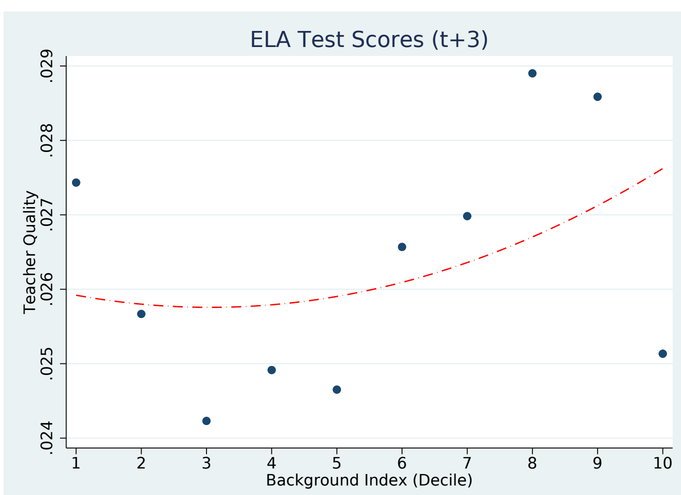
Figure 3: Differences in Teacher Quality by Background Index: English



(a) Contemporaneous test scores



(b) Test scores at $t + 2$



(c) Test scores at $t + 3$

Table 4: Variance Decomposition: Pooled Subjects

	Outcomes		
	Test Score t	Test Score $t+2$	Test Score $t+3$
Variance component	(1)	(2)	(3)
Total: $Var(y_{ijsgt})$	0.958	0.953	0.941
Background: $Var(X'_{ijsgt}\beta)$	0.657	0.568	0.546
School: $Var(\phi_s)$	0.010	0.017	0.022
Teacher: $Var(\tau_{jt})$	0.023	0.021	0.023
Background-School Covariance: $2 \cdot Cov(X'_{ijsgt}, \phi_s)$	0.013	0.023	0.030
Background-Teacher Covariance: $2 \cdot Cov(X'_{ijsgt}, \tau_{jt})$	0.010	0.004	0.005
School-Teacher Covariance: $2 \cdot Cov(\phi_s, \tau_{jt})$	-0.014	-0.022	-0.024
Error term: $Var(\epsilon_{ijsgt})$	0.259	0.323	0.337
Schools:	498	498	498
Observations:	1,531,198	1,346,988	1,281,552

Notes: Each decomposition is based on the estimates from (3.1) using pooled data for fourth-grade and fifth-grade students taking one of the end-of-grade ELA or math tests. School quality (ϕ_s) corresponds to the estimated school fixed effect. Teacher quality corresponds to the sum of the teacher fixed effect and the estimate of experience: $\tau_{jt} = \hat{\mu}_j + \hat{f}(exp_{jt}) + \hat{f}^S(exp_{jst})$.

Table 5: Variance Decomposition: Math

	Outcomes		
	Test Score t	Test Score $t+2$	Test Score $t+3$
Variance component	(1)	(2)	(3)
Total: $Var(y_{ijsgt})$	0.963	0.963	0.963
Background: $Var(X'_{ijsgt}\beta)$	0.670	0.622	0.571
School: $Var(\phi_s)$	0.015	0.023	0.031
Teacher: $Var(\tau_{jt})$	0.040	0.025	0.032
Background-School Covariance: $2 \cdot Cov(X'_{ijsgt}, \phi_s)$	0.121	0.026	0.033
Background-Teacher Covariance: $2 \cdot Cov(X'_{ijsgt}, \tau_{jt})$	0.014	0.015	0.010
School-Teacher Covariance: $2 \cdot Cov(\phi_s, \tau_{jt})$	-0.022	-0.024	-0.032
Error term: $Var(\epsilon_{ijsgt})$	0.235	0.299	0.325
Schools:	474	474	474
Observations:	774,401	673,449	638,903

Notes: Each decomposition is based on the estimates from (3.1). School quality (ϕ_s) corresponds to the estimated school fixed effect. Teacher quality corresponds to the sum of the teacher fixed effect and the estimate of experience: $\tau_{jt} = \hat{\mu}_j + \hat{f}(exp_{jt}) + \hat{f}^S(exp_{jst})$.

Table 6: Variance Decomposition: English

	Outcomes		
	Test Score t	Test Score $t+2$	Test Score $t+3$
Variance component	(1)	(2)	(3)
Total: $Var(y_{ijsgt})$	0.953	0.953	0.953
Background: $Var(X'_{ijsgt}\beta)$	0.642	0.568	0.521
School: $Var(\phi_s)$	0.012	0.017	0.022
Teacher: $Var(\tau_{jt})$	0.021	0.021	0.026
Background-School Covariance: $2 \cdot Cov(X'_{ijsgt}, \phi_s)$	0.015	0.023	0.028
Background-Teacher Covariance: $2 \cdot Cov(X'_{ijsgt}, \tau_{jt})$	0.006	0.004	0.000
School-Teacher Covariance: $2 \cdot Cov(\phi_s, \tau_{jt})$	-0.018	-0.022	-0.028
Error term: $Var(\epsilon_{ijsgt})$	0.275	0.323	0.343
Schools:	471	471	471
Observations:	765,789	641,535	612,337

Notes: Each decomposition is based on the estimates from (3.1). School quality (ϕ_s) corresponds to the estimated school fixed effect. Teacher quality corresponds to the sum of the teacher fixed effect and the estimate of experience: $\tau_{jt} = \hat{\mu}_j + \hat{f}(exp_{jt}) + \hat{f}^S(exp_{jst})$.

Table 7: Between-School and Within-School Differences in Teacher Quality: Pooled Subjects

		Average Differences in Teacher Quality by Background Score		
	Test Score (y_i)	Within-School ($\hat{\tau}_{jt} - \bar{\hat{\tau}}_s$)	Between-School ($\bar{\hat{\tau}}_s$)	Total ($\hat{\tau}_{jt}$)
Panel A: Test Score at t				
Top 10% - Bottom 10%	2.811	0.017	0.004	0.021
Top 25% - Bottom 25%	2.153	0.012	0.002	0.014
Panel B: Test Score at $t + 2$				
Top 10% - Bottom 10%	2.687	0.012	0.007	0.018
Top 25% - Bottom 25%	2.075	0.009	0.006	0.015
Panel C: Test Score at $t + 3$				
Top 10% - Bottom 10%	2.596	0.011	-0.001	0.010
Top 25% - Bottom 25%	1.993	0.009	0.001	0.010

Notes: For each panel, each cell corresponds to the estimated difference between the top and bottom quartiles and top and bottom deciles of the background score distribution. The background score is computed for each student using the estimated value $X'_{ijgst}\hat{\beta}$. Total teacher quality corresponds to the sum of the teacher fixed effect and the estimates of experience: $\hat{\tau}_{jt} = \hat{\mu}_j + \hat{f}(exp_{jt}) + \hat{f}^S(exp_{jst})$. The school average $\bar{\hat{\tau}}_s$ corresponds to the average value of $\hat{\tau}_{jt}$ for teachers observed in school s, weighted by the number of students taught by each teacher.

Table 8: Bias-Corrected Variance Decomposition: Pooled Subjects

	Outcomes		
	Test Score t	Test Score $t+2$	Test Score $t+3$
Variance component	(1)	(2)	(3)
Total: $Var(y_{ijsgt})$	0.956	0.959	0.939
School: $Var(\phi_s)$	0.007	0.013	0.018
Teacher: $Var(\tau_{jt})$	0.018	0.014	0.020
School-Teacher Covariance: $2 \cdot Cov(\phi_s, \tau_{jt})$	-0.004	-0.005	-0.009
Schools:	480	480	480
Observations:	1,490,091	1,310,267	1,246,457

Notes: Each decomposition is based on the estimates from (3.1). School quality (ϕ_s) corresponds to the estimated school fixed effect. Teacher quality corresponds to the sum of the teacher fixed effect and the estimate of experience: $\tau_{jt} = \hat{\mu}_j + \hat{f}(exp_{jt}) + \hat{f}^S(exp_{jst})$.

References

J. Abowd, F. Kramarz, and D. Margolis. "High Wage Workers and High Wage Firms". *Econometrica*, 67(2):251–334, 1999.

S. Bonhomme, K. Holzheu, T. Lamadon, E. Manresa, M. Mogstad, and B. Setzler. "How Much Should we Trust Estimates of Firm Effects and Worker Sorting?". NBER Working Paper 27368, 2020.

D. Card, J. Heining, and P. Kline. "Workplace Heterogeneity and the Rise of West German Wage Inequality". *The Quarterly Journal of Economics*, 128(3):967–1015, 2013.

D. Card, A. R. Cardoso, and P. Kline. "Bargaining, Sorting, and the Gender Wage Gap: Quantifying the Impact of Firms on the Relative Pay of Women". *The Quarterly Journal of Economics*, 131(2):633–686, 2015.

R. Chetty, J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star". *The Quarterly Journal of Economics*, 126(4):1593–1660, 2011.

C. Clotfelter, H. Ladd, and J. Vigdor. "Who Teaches Whom? Race and the Distribution of Novice Teachers". *Economics of Education Review*, 24(4):377–392, 2005.

M. Gilraine and N. Pope. "Making Teaching Last: Long- and Short-Run Value Added". Working paper, 2020.

D. Goldhaber, L. Lavery, and R. Theobald. "Uneven Playing Field? Assessing the Teacher Quality Gap Between Advantaged and Disadvantaged Students". *Educational Researcher*, 44(5):293–307, 2015.

E. A. Hanushek, J. F. Kain, and S. G. Rivkin. "Why Public Schools Lose Teachers". *Journal of Human Resources*, 39(2):326–354, 2004.

C. K. Jackson. "Student Demographics, Teacher Sorting, and Teacher Quality: Evidence from the End of School Desegregation". *Journal of Labor Economics*, 27(2):213–256, 2009.

C. K. Jackson. "Match Quality, Worker Productivity, and Worker Mobility: Direct Evidence from Teachers". *The Review of Economics and Statistics*, 95(4):1096–1116, 2013.

C. K. Jackson. "Teacher Quality at the High School Level: The Importance of Accounting for Tracks". *Journal of Labor Economics*, 32(4):645–684, 2014.

C. K. Jackson. "What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes". *Journal of Political Economy*, 126(5):2072–2107, 2018.

C. K. Jackson, J. E. Rockoff, and D. O. Staiger. "Teacher Effects and Teacher-Related Policies". *Annual Review of Economics*, 6(1):801–825, 2014.

B. A. Jacob, L. Lefgren, and D. P. Sims. “The Persistence of Teacher-Induced Learning”. *Journal of Human Resources*, 45(4):915–943, 2010.

K. Jochmans and M. Weidner. “Fixed-Effect Regressions on Network Data”. *Econometrica*, 87(5): 1543–1560, 2019.

P. Kline, R. Saggio, and M. Sølvsten. “Leave-Out Estimation of Variance Components”. *Econometrica*, 88(5):1859–1898, 2020.

H. Lankford, S. Loeb, and J. Wyckoff. “Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis”. *Educational Evaluation and Policy Analysis*, 24(1):37–62, 2002.

R. Mansfield. “Teacher Quality and Student Inequality”. *Journal of Labor Economics*, 33(3): 751–788, 2015.

N. Petek and N. Pope. “The Multidimensional Impact of Teachers on Student Outcomes”. Working Paper, Department of Economics, University of Maryland, 2021.

S. F. Reardon, D. Kalogrides, and K. Shores. “The Geography of Racial/Ethnic Test Score Gaps”. *American Journal of Sociology*, 124(4):1164–1221, 2019.

T. R. Sass, J. Hannaway, Z. Xu, D. N. Figlio, and L. Feng. “Value Added of Teachers in High-Poverty Schools and Lower Poverty Schools”. *Journal of Urban Economics*, 72(2):104–122, 2012.

P. Thiemann. “Inequality in Education Outcomes: The Role of Sorting among Students, Teachers, and Schools”. Working Paper, 2019.