Inequality in Education Outcomes: The Role of Sorting among Students, Teachers, and Schools*

Petra Thiemann†[1]

[1]Lund University and IZA

October 16, 2018

This paper studies the contribution of non-random sorting ("assortative matching") among students, teachers, and schools to test score inequality across classrooms, accounting for both observed and unobserved teacher and school inputs. I recover teacher and school unobservables based on teacher switching across schools using longitudinal data from the universe of public elementary schools in North Carolina. Assortative matching accounts for a substantial portion of the across-classroom variance in student achievement (17-19%). One-third of this contribution is due to teacher-classroom sorting within and across schools, and two-thirds are due to student-school sorting. This pattern is driven by urban and suburban areas.

JEL codes: I20, I24, J24
Keywords: education, inequality, K-12, teacher quality, school quality, teacher switching, student sorting, assortative matching

*An earlier version of the paper was circulated under the title "Does Teacher Sorting within and across schools contribute to Inequality in Education Outcomes?".

†Contact: Lund University, Department of Economics, School of Economics and Management, P.O. Box 7082, SE-220 07 Lund, Sweden, Phone: +46 (0)46 222 8686, E-mail: petra.thiemann@nek.lu.se.
Petra Thiemann was a Postdoctoral Research Associate at USC Dornsife INET while working on this paper and is currently still affiliated with INET. The paper greatly benefited from feedback by Jan Bietenbeck, Khai X. Chiong, Bryan S. Graham, Kirabo Jackson, Matthew Kahn, Michael Leung, Kaveh Majlesi, Elena Manresa, Alessandro Martinello, Hyungsik Roger Moon, Amine Ouazad, Geert Ridder, Jorge De La Roca, Jorge Tamayo, Valentin Verdier, Jenny Williams, as well as seminar and conference participants. I thank Young Miller from the USC Economics Department, Kara Bonneau from the North Carolina Education Research Data Center at Duke, and Jeremy Holt from USC Dornsife IT Services for their generous administrative support. I acknowledge funding from the USC Dornsife Institute for Economic Thinking and the USC Department of Economics.

1 Introduction

Teachers and schools are key determinants of educational success and later-life outcomes (Rivkin et al., 2005, Chetty et al., 2014a,b, Dobbie and Fryer, 2013, Angrist et al., 2016). However, not all students have equal access to high-quality teachers and schools. For example, students from minority or disadvantaged backgrounds are on average matched with teachers with lower levels of experience, preparation, or value-added (Clotfelter et al., 2005, 2007, Lankford et al., 2002, Sass et al., 2012). Similarly, good teachers are often assigned to high-achieving classrooms (Dieterle et al., 2015).[1] Differences in access to high-quality teachers and high-quality schools, in turn, can exacerbate inequalities in education outcomes (Reardon et al., 2016). Little is known, however, about the extent to which unequal access to teacher and school inputs contributes to inequality in education outcomes. So far, the literature on this topic remains inconclusive.[2]

In this paper, I provide new evidence on the question of whether unequal access to teacher and school inputs explains inequality in education outcomes. I start from the observation that students and teachers are assigned to (or self-select into) schools and classrooms; these assignments or "sorting" processes lead to a pattern where high-ability students attend good schools and learn from good teachers more frequently than low-ability students. In other words, the resulting student-teacher-school-matches are "assortative" (i.e., non-random), which in turn can drive inequality in education outcomes. My analysis disentangles the relative contribution of the three different components of assortative matching—the matching of teachers to schools, the matching of students to schools, and the matching of teachers to students. In this way, I explain variations in test score outcomes among elementary school students in North Carolina.

The methodological challenge of the analysis lies in the identification and estimation of

[1] Teacher sorting accounts for some of these matching patterns. Teachers prefer schools with favorable student backgrounds and sort on the ability of the students (Boyd et al., 2013). Moreover, teachers have preferences over the demographic composition of the student body in a school (Jackson, 2009), and good teachers tend to leave schools that do not meet accountability standards (Feng et al., 2018).

[2] Mansfield (2015) concludes that teacher-student sorting explains only a negligible share in the test score variance across students in North Carolina's public high schools. Reardon et al. (2016) highlight that racial and ethnic segregation across schools is important in explaining achievement gaps across students in several metropolitan areas and school districts in the US.

teacher and school quality since observable teacher and school characteristics are in general poor predictors for student achievement.[3] To address this challenge, this paper uses identification results from the worker-firm sorting literature (Abowd et al., 1999, henceforth "AKM") to disentangle and estimate the unobserved components of teacher and school quality in an education production model. In the worker-firm context, the switching of workers across firms generates the identifying variation which allows the researcher to disentangle the unobserved components of a firm's effectiveness from the unobserved component of a worker's productivity. Similarly, in the teacher-school setting, the switching of teachers across schools generates such identifying variation.[4] The education setting, however, is different because of student sorting, which adds an additional layer to the sorting problem. Students select into schools based on teacher and school characteristics, and teachers select into schools based on the composition of the student body. I extend the AKM-type education production model such that students can sort based on observed and unobserved characteristics of the teachers and schools, and teachers can sort based on observed student characteristics, in particular, past test scores and demographic characteristics.

The data set covers the universe of public elementary schools in North Carolina; it contains about 1.6 million student-year observations and extends over a period of 15 years (1997-2011).[5] 20 percent of the teachers switch schools at least once during the period under study and thus generate the identifying variation. The outcomes are standardized end-of-year test scores in math and reading. In order to measure students' preparedness, I use standardized pretest scores as well as a rich set of variables on students' socio-economic backgrounds.

I find that the assortative matching between students, schools, and teachers explains in total 17 percent of the across-classroom test score variation in math, and 19 percent of the

[3] Observable teacher characteristics such as experience and education explain only a small portion of a teacher's effectiveness (Rivkin et al., 2005, Rockoff, 2004). Similarly, observable school characteristics usually explain only a small portion of school effectiveness (Dobbie and Fryer, 2013, Hanushek, 1997). Recent papers thus measure school and teacher effectiveness based on unobservable characteristics using longitudinal data (Rockoff, 2004, Chetty et al., 2014a, Rothstein, 2017).

[4] An alternative method would be to exploit student switching across schools to recover student unobservables. This approach is taken in a related work by Kramarz et al. (2015).

[5] Similar data sets have been constructed based on the NC data by Jackson (2013) to study match effects between teachers and schools and by Rothstein (2017) to study biases in value-added estimates.

across-classroom variation in reading.[6] About one-third of this variation is due to teacher sorting both within and across schools, and the other two-thirds are due to student-school sorting. In a simulation exercise, I show that removing inequality in teacher access diminishes the test score gap between the highest and lowest achieving classrooms. The test score gap between classrooms at the top and bottom decile of the test score distribution would shrink by 13 percent (0.16 test score standard deviations) in math and 8 percent (0.10 test score standard deviations) in reading if teachers were randomly distributed across classrooms and schools, holding classroom composition and all other factors constant. Finally, the model uncovers systematic heterogeneity in sorting patterns across geographic locations. In large cities, the sorting of students to schools explains a larger portion of the variation than the sorting of teachers to students (19 percent versus 3 percent in math; 21 percent versus 4 percent in reading). In rural areas, the reverse is the case, i.e. the sorting of students to schools explains a smaller portion of the variation than the sorting of teachers to students (3 versus 10 percent in math and 2 versus 7 percent in reading).

This paper relates to a number of studies that investigate unequal access to school and teacher resources among students from different backgrounds. One strand of the literature documents patterns of residential sorting and school segregation, which negatively affect disadvantaged students (e.g. Deming et al., 2014a, Owens et al., 2016, Reardon et al., 2016).[7] A second strand of the literature analyzes teacher sorting within and across schools and shows that disadvantaged students often have access to low-quality teachers, both within schools (Boyd et al., 2006, Dieterle et al., 2015, Kalogrides and Loeb, 2013), and across schools (Clotfelter et al., 2005, 2007, Lankford et al., 2002, Jackson, 2009, Feng, 2010, Sass et al., 2012, Feng et al., 2018). This literature, however, does not directly investigate the implications of such sorting patterns for inequality in education outcomes. An exception is an article by Mansfield (2015), who studies the link between student-teacher sorting in North Carolina's high schools and inequality in student test score outcomes. He finds that teacher

[6]The remaining variation is explained by student background characteristics (52% of the variation in math and 50% of the variation in reading) and (non-systematic) variation in teacher quality (16% in math and 9% in reading) and school quality (7% in math and 6% in reading).

[7]This literature also emphasizes the importance of accounting for unobservable school inputs when measuring school quality (Deming et al., 2014b).

sorting explains only a negligible share of the variation in student test scores.

In the present paper, I contribute to the literature by analyzing three different sorting channels (teacher-school, teacher-student, and student-school) and their relative importance in a single model. This allows me to detect and compare different sources of inequality in the education system, rather than focusing on a single channel. I show that all sorting channels combined contribute substantially to the variation in test score outcomes in the context of North Carolina's primary schools; most of this variation is driven by student-school sorting, but teacher-classroom sorting also plays a non-negligible role. These results complement prior analyses that show only small contributions of teacher sorting at higher levels of schooling (Mansfield, 2015). In addition, the present paper uncovers sizable heterogeneity in sorting patterns across different geographic locations, thus emphasizing the need to take spatial differences into account when explaining inequality in student achievement and later-life outcomes.

The paper proceeds as follows. Section 2 presents a stylized overview of the sorting problem as well as the institutional background for teacher assignments in North Carolina. Section 3 outlines the empirical approach, and Section 4 presents the data sets and variables. Section 5 describes and discusses the results. Section 6 discusses the findings in the light of the literature, and Section 7 concludes.

2 Background

This section motivates and informs the empirical model. First, I review the institutional background that governs teacher and student sorting in North Carolina's public elementary school system. Second, I discuss expected sorting patterns based on empirical findings from North Carolina and other states.

Institutional background. The teacher assignment to schools and students in North Carolina's public elementary schools proceeds in two main steps. In the first step, teachers form a job match with a school. Teachers either apply directly at the school, or they apply to the school district, and the school district refers them to a school (Jackson, 2009). Teachers can be both new teachers as well as incoming teachers from a different school. In the second step, at the beginning of each school year, school principals assign teachers to classrooms. This paper focuses on the main classroom teachers in math and reading, who teach only one classroom in a given year.

Because of the regulation of teacher pay, limited variations of teacher wages across and within schools exist, which restricts the scope for schools to attract teachers through monetary incentives. A state-wide pay schedule allows teacher pay to vary only with experience and education. School districts can pay a supplement to the state salary, but this supplement can only vary by experience and education as well (Clotfelter et al., 2011).

The student sorting consists of two main steps as well. In the first step, students sort to schools. In North Carolina, a student's residential location largely determines his/her assignment to a public school, with few exceptions (Bifulco and Ladd, 2007, Bifulco et al., 2009, Jackson, 2009); but students can opt out of the public school system and choose private schools, charter schools, or homeschooling instead. In the second step, principals assign students to classrooms and thus match them with a teacher.

Expected sorting patterns. From the perspective of standard labor supply theory, one might not expect assortative matching between teachers and students/schools in settings where wage differentiation across teachers is very limited. Several studies, however, document assortative matching among students, teachers, and schools in such settings. Better-prepared students have access to better teachers, both within and across schools (Dieterle et al., 2015, Clotfelter et al., 2006, Lankford et al., 2002, Sass et al., 2012), better teachers teach in better schools (Feng et al., 2018), and schools with favorable student characteristics attract better teachers (Sass et al., 2012, Boyd et al., 2013).

Teachers often sort based on the observed characteristics of the students in a school (Clotfelter et al., 2011, Boyd et al., 2013). Teachers generally favor schools and classrooms with higher ability levels; moreover, teachers have preferences over students' demographic characteristics—their racial and socio-economic background, for example, (Dieterle et al., 2015, Jackson, 2009, Sass et al., 2012). The reasons for these preferences are diverse; for example, high-ability students might be easier to teach, and schools with lower poverty levels are often in more attractive neighborhoods. In addition, school principals use classroom assignments to attract, reward, or retain good teachers. Thus, within a school, better teachers are likely to teach better-prepared students (Dieterle et al., 2015).[8]

Teacher sorting patterns and student sorting patterns are intertwined and likely to reinforce each other. While a teacher may choose a school based on the demographic composition of the student body (Boyd et al., 2013), families may choose their residential location based on the composition of the teacher work force in the desired school or school district.

The sorting between students, schools, and teachers, however, is unlikely to turn out as perfect positive assortative matching for a number of reasons. First, within schools, principals may want to reduce the inequality in test score outcomes across students. In this case, they may assign better teachers to disadvantaged students in order to narrow the achievement gap. Second, teachers' preferences over students' characteristics are heterogeneous. While some of the teachers prefer classrooms that are easy to teach, others prefer to teach classrooms with low-ability or disadvantaged students. Third, location constraints, search costs, and job-switching costs add additional frictions, which have been well-studied in the labor literature (Mortensen, 1986, Mortensen and Pissarides, 1999). For example, a good teacher may prefer to teach in a high-quality school, but may not be willing to incur commuting or relocation costs (Boyd et al., 2013).

Ultimately, the amount of assortativeness in the data is an open empirical question; therefore, the next section turns to the empirical model, which quantifies the amount of sorting and its relation to test score inequality across classrooms.

[8] Alternative explanations for sorting are complementarities between characteristics of the teacher and the classroom. This paper, however, does not model such complementarities in order to keep the model tractable.

3 Empirical approach

3.1 Model

I consider an educational production function, where the output variables are end-of-grade test scores of students in grades three to five in math and reading. This output depends on student inputs, teacher inputs, and school inputs:

$$y_{it} = \underbrace{X'_{it}\gamma}_{\text{student preparedness}} + \underbrace{\mu_{J(i,t)} + V'_{J(i,t)t}\delta}_{\text{teacher quality}} + \underbrace{\alpha_{S(i,t)} + W'_{S(i,t)t}\rho}_{\text{school quality}} + \underbrace{D'_t\psi + G'_{it}\xi}_{\text{grade and year dummies}} + \underbrace{\epsilon_{it}}_{\text{error term}} . \quad (1)$$

The outcome y_{it} is the test score of student i at time t. Three types of inputs enter into the production of test scores in an additively separable way. First, the outcome depends on a student's background X_{it}, which is a vector of characteristics including measures of ability (baseline test scores before entering the third grade, classified as gifted or learning disadvantaged, limited English proficiency), socio-economic background (free/reduced-price lunch eligibility, parental education), and demographic characteristics (gender, race, age). Second, the model includes teacher quality. Teacher quality has a time-invariant component also known as "value-added", $\mu_{J(i,t)}$, where $J(i,t) = j$ is a function that uniquely maps student i to its teacher j at time t; teacher quality, furthermore, has a time-varying component measured through teacher experience, $V_{J(i,t)t}$.[9] Third, independent of teacher quality, the school provides several inputs such as the school administration and facilities. The time-invariant component of school quality is denoted as $\alpha_{S(i,t)}$, where $S(i,t) = s$ is the function that maps the student i to its school s at time t. Moreover, the school quality may vary with the composition of the student body in terms of background characteristics, $W_{S(i,t)t}$ (average baseline test scores at the school level, the fraction of students with free/reduced-price lunch, and the racial composition). Finally, test scores can vary across years, captured by year dummies, D_t, as well as across grades, captured by grade dummies, G_t. ϵ_{ijst} is an idiosyncratic error term.

[9] In order to separately identify time effects from experience effects, I pool experience levels into categories (c.f. Mansfield, 2015): less than one year, 1-2 years, 3-5 years, 6-11 years, 12 and more years.

The error term consists of three different components:

$$\epsilon_{it} = \phi_{S(i,t)t} + \nu_{J(i,t)t} + e_{it}, \tag{2}$$

where $\phi_{S(i,t)t}$ captures shocks to the impact of school quality (e.g., through a change of the school principal), $\nu_{J(i,t)t}$ captures shocks to teacher value-added (e.g., through changes in the teacher's health), and e_{it} captures all other shocks to student outcomes (e.g., shocks to parental inputs).

This paper quantifies the contribution of teacher quality, school quality, and assortative matching between teachers, schools, and students to the dispersion in test scores across students; to this end, I decompose the test score variance of test scores into several components. The variance of the sum on the right-hand side of Equation 1 can be rewritten as the sum of the variances plus all possible covariances. Following Abowd et al. (1999), I exploit this result in order to decompose the variance into its main components (variances) and the sorting components (covariances).

The following simplified example of an education production function illustrates the variance decomposition and its interpretation. Consider a random variable Y, here, student test scores. Assume further that the test scores are solely determined by two other random variables, student preparedness X_1 and teacher quality X_2, and that both inputs are additively separable such that $Y = X_1 + X_2$. Then, we can decompose the test score variance as

$$\text{Var}(Y) = \text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) + 2\text{Cov}(X_1, X_2). \tag{3}$$

The first two terms capture the variances in student and teacher quality. The last term, $2\text{Cov}(X_1, X_2)$, captures both the amount of sorting and its influence on the outcome. For example, if $\text{Cov}(X_1, X_2) = 0$, then the teacher and student inputs are uncorrelated, and sorting does not contribute to differences in test scores across students. One way to achieve a zero covariance is, for example, to assign students randomly to teachers. Notice, however, that random assignment is not a necessary condition for a zero covariance. For example, if teachers and students sort along dimensions that are unrelated to their preparedness/quality, the sorting does not contribute to inequality in test score outcomes. If $\text{Cov}(X_1, X_2) > 0$, then positive assortative matching exists, and sorting exacerbates the test score differences across students. This, for example, is the case if better teachers are systematically assigned to better-prepared students. If $\text{Cov}(X_1, X_2) < 0$, then negative assortative matching exists, and sorting reduces the test score gap between highly-prepared students and students with low levels of preparation. This, for example, is the case if students with low levels of preparation are systematically assigned to high-quality teachers.

Model 1 is more complex and contains a larger number of variables, yet the same principle applies—the variances capture the contribution of the inputs, and the covariances measure the contribution of sorting. The model distinguishes between the three main input categories, student preparedness, teacher quality (a combination of teacher value-added and teacher experience), and school quality (school effects and school composition); thus, the decomposition results in three variance and three covariance terms, i.e. six parameters of interest (see the definition of parameters in Table A.1).

A focus of this paper is on the sorting of better teachers to better schools and students; the following term captures the sorting of high-quality teachers to highly-prepared students:

$$\text{Cov}((\mu_{J(i,t)} + V'_{J(i,t)t}\delta), X'_{it}\gamma). \tag{4}$$

Moreover, the decomposition is informative about the sorting of students to schools. For example, the following term captures the sorting of highly-prepared students to high-quality schools:

$$\text{Cov}((\alpha_{S(i,t)} + W'_{S(i,t)t}\rho), X'_{it}\gamma). \tag{5}$$

After decomposing the variance into its parts, one can compare the different sorting processes and express the variance contributions as a fraction of the total variance. For example, the portion of the variance in test scores that can be explained by the sorting of highly-prepared students to high-quality teachers is:

$$\frac{\text{Cov}((\mu_{J(i,t)} + V'_{J(i,t)t}\delta), X'_{it}\gamma)}{\text{Var}(y_{ijst})}. \tag{6}$$

The analysis, thus, informs researchers and policy makers about the relative importance of different sorting channels in a given setting. For example, are school and teacher resources relatively balanced across students? Additionally, the model allows for a comparison of these patterns across time and space; this paper considers heterogeneity in the sorting patterns across different geographical areas (e.g., urban and rural areas).

3.2 Identification

Model 1 is a version of the worker-firm sorting model by Abowd et al. (1999) (also abbreviated as "AKM" model). In their model, the authors study the sorting of workers to firms/establishments along unobservable dimensions. This is similar to the setting in this paper, where the schools are the "firms", and the teachers are the "workers". The AKM model and Model 1 differ mainly because Model 1 measures the outcome at the student and not at the teacher (worker) level. The sorting of students to schools and teachers adds an additional layer to the model.

In order to discuss the assumptions for the model to be identified, I follow Abowd et al. (1999) and Card et al. (2013) and rewrite the model in matrix notation. N^* denotes the number of student-year observations.[10] I furthermore introduce J teacher indicators, and S school indicators. These indicators are set to 1 if the student is assigned to the respective teacher/school, and 0 otherwise. Each student can only be assigned to one school and teacher in a given year. Thus, Model 1 can be rewritten as:

$$y = \underbrace{X\gamma}_{\text{student preparedness}} + \underbrace{H\mu + V\delta}_{\text{teacher quality}} + \underbrace{F\alpha + W\rho}_{\text{school quality}} + \underbrace{D\psi + G\xi}_{\text{year and grade dummies}} + \epsilon, \tag{7}$$

where y is an $N^* \times 1$ vector of test scores, $H \equiv [h^1, ..., h^J]$ is an $N^* \times J$ matrix of teacher

[10] Each student is present in the data set for at most three years, with the exception of students who repeat a year.

indicators, and $F \equiv [f^1, ..., f^S]$ is an $N^* \times S$ matrix of school indicators. The model also contains a matrix of student-level controls $X \equiv [x^1, ..., x^K]$, a matrix of indicators for teachers' experience levels $V \equiv [v^1, ..., v^L]$,[11] a matrix of school composition variables $W \equiv [w^1, ..., w^M]$, a matrix of time dummies, $D \equiv [d^1, ..., d^T]$, and a matrix of grade dummies $G \equiv [g^1, ..., g^P]$. The matrices for the dummy variables are always defined up to one reference category.

The first set of assumptions ensures that teacher and school effects can be separately identified. First, schools must be observed multiple times during the sample period, and at least some of the teachers must be observed in different schools. This implies that some of the teachers must switch schools. To illustrate the importance of this assumption, suppose that each school hires a fixed set of teachers in the first time period and remains with these teachers throughout the whole sample period. The school effect would then be completely absorbed by the teacher effects.

Second, the identification requires that a set of schools is completely connected through teacher switches during the sample period. Two schools form a link if at least one teacher switches between the two schools. All schools in the sample must be linked directly or indirectly to one another because the school effects can only be interpreted relative to a reference school. In a connected set of schools, the school effects are identified up to one reference school. The number of reference schools equals the number of connected sets. If several connected sets are pooled in the analysis, one cannot directly compare the school effects without further assumptions.

A second set of assumptions considers the strict exogeneity of the observed characteristics, the teacher assignments, and the school assignments. Conditional on all variables in the model, each of the variables and assignments must be uncorrelated with contemporaneous, past, or future shocks to the outcomes. Given the seven matrices on the right hand side of Equation 7, we express this assumption in terms of eight orthogonality conditions:

[11] Experience levels are broken up into bins (less than one year, 1-2 years, 3-5 years, 6-11 years, 12 years and more) so that they can be identified separately from year fixed effects.

$$E[x^{k'}\epsilon] = 0 \quad \forall k, \quad E[h^{j'}\epsilon] = 0 \quad \forall j, \quad E[v^{l'}\epsilon] = 0 \quad \forall l, \tag{8}$$
$$E[f^{s'}\epsilon] = 0 \quad \forall s, \quad E[w^{m'}\epsilon] = 0 \quad \forall m, \quad E[d^{t'}\epsilon] = 0 \quad \forall t, \quad E[g^{p'}\epsilon] = 0 \quad \forall p.$$

Three of the orthogonality conditions concern the set of student background characteristics as well as time and grade dummies:

$$E[x^{k'}\epsilon] = 0 \quad \forall k, \quad E[d^{t'}\epsilon] = 0 \quad \forall t, \quad E[g^{p'}\epsilon] = 0 \quad \forall p. \tag{9}$$

I assume that student background characteristics and year and grade indicators are predetermined. In other words, a student's background (e.g., his baseline test score or demographic characteristics) must not be affected by contemporaneous or future shocks to the outcome, conditional on all other characteristics in the model. This assumption is standard in the literature since it is difficult to manipulate or adjust background characteristics; moreover, the baseline test scores and demographic characteristics are measured before the child enters school.

Two further orthogonality conditions consider the teacher assignment:

$$E[h^{j'}\epsilon] = 0 \quad \forall j, \tag{10}$$
$$E[v^{l'}\epsilon] = 0 \quad \forall l. \tag{11}$$

As Card et al. (2013) show in the worker-firm context, a sufficient condition for Equation 10 to hold is that teacher assignments are independent of contemporaneous, past, and future shocks to student achievement, teacher quality, and school quality, conditional on all other variables in the model. In the formulation of Model 1, this implies that

$$\begin{aligned} P[J(i,t) &= j|\boldsymbol{\epsilon}, \alpha_{S(i,t)}, \mu_{J(i,t)}, X_{it}, V_{J(i,t)t)}, W_{S(i,t)t}, D_t, G_t] \\ = P[J(i,t) &= j|\alpha_{S(i,t)}, \mu_{J(i,t)}, X_{it}, V_{J(i,t)t)}, W_{S(i,t)t}, D_t, G_t] \; \forall \; i, j, t, s. \end{aligned} \tag{12}$$

The *probability* of being assigned to a certain teacher must thus be independent of contemporaneous, past, or future shocks to teacher quality, school quality, or any other idiosyncratic shocks to the outcome. This assumption is not violated if better-prepared students are matched with better teachers. As long as student background characteristics sufficiently reflect student preparedness (e.g., through prior test scores and other socio-demographic characteristics), sorting does not bias the model estimates.

There are at least four concerns about the strict exogeneity of teacher assignments (Equation 12):

First, this assumption is violated if teachers sort into schools based on anticipated shocks to school inputs. For example, suppose that a teacher switches to a school that he anticipates will become better after the switch (for reasons other than the quality of the other teachers at the school). In this case, the model would attribute the quality increase to the teacher fixed effect, rather than to the school fixed effect, so that the teacher fixed effect will be upward biased. In order to mitigate this concern, the model includes time-varying school characteristics that reflect changes in school quality (i.e., average test scores at the school level, racial and socio-economic composition of the student body).

Second, the assumption is violated if teacher effort displays an Ashenfelter dip (or spike), i.e. if teachers slack off once they know that they will change their workplace, or they spend disproportionately high effort once they start working in a new school. In this case, the model would wrongly attribute the adjustments in effort to the quality of the school. This would bias some of the school fixed effects upwards, and some of them downwards. If this bias is systematically related to school quality (e.g., teachers tend to increase their effort more when changing to a higher-quality school), one would overstate the importance of school fixed effects. Since both effort and school quality are unobserved in the present setting, however, it is not possible to directly test this assumption.

Third, the assumption can be violated through dynamic tracking or sorting of students on unobservable student characteristics, even conditional on lagged student test scores (Rothstein, 2010). For example, teachers who are going to leave a school may be systematically assigned to students that are worse in unobserved dimensions, and teachers who are new to a school may be systematically assigned to students that are better in unobserved dimensions. In this case, one would attribute the effect of student ability to the permanent teacher effect, which could inflate the variance in teacher value-added (Rothstein, 2017). Other sources of dynamic sorting are time-varying student characteristics that are not included in the model. Controlling for a large set of student background characteristics is, therefore, essential in the present setting.

Fourth, the assumption may be violated if match effects or complementarities between teachers and schools occur. Specifically, teachers who switch may be more productive in their new school, compared to their old school, as they learn about their school-specific productivity over time (Jackson, 2013). In this case, one can no longer identify the school fixed effects separately from the teacher effects. If teachers improve their match quality over time, the variables that capture teacher experience absorb part of the match effects. To test for remaining teacher-school match effects, I follow Card et al. (2013) and compare the model fit of the AKM-type model with the fit of a model that uses fully interacted teacher-by-school fixed effects. Since the model fit of the fully interacted model is only slightly better than the model fit of the AKM-type model, I conclude that the match effects are unlikely to affect the aggregate results on the contribution of sorting (see Section A.3.1 for details).

I also assume that the experience level of the assigned teacher is pre-determined, so that Equation 11 holds. Finally, the last two orthogonality conditions consider the assignment of students to schools:

$$E[f^{s'}\epsilon] = 0 \quad \forall s, \tag{13}$$

$$E[w^{m'}\epsilon] = 0 \quad \forall m. \tag{14}$$

In order for Equation 13 to hold, school assignment has to fulfill the following strict exogeneity condition, expressed as the probability of student i to be assigned to a certain school in a given year:

$$\begin{aligned} P[S(i,t) &= s|\boldsymbol{\epsilon}, \alpha_{S(i,t)}, \mu_{J(i,t)}, X_{it}, V_{J(i,t)t}, W_{S(i,t)t}, D_t, G_t] \qquad (15) \\ = P[S(i,t) &= s|\alpha_{S(i,t)}, \mu_{J(i,t)}, X_{it}, V_{J(i,t)t)}, W_{S(i,t)t}, D_t, G_t] \; \forall \; i, j, s, t. \end{aligned}$$

The probability distribution over school choices for a student in a given year must be independent of past, present, or future shocks to teacher value-added, school value-added, or other idiosyncratic shocks to student outcomes, conditional on all other variables included in the model. The assumption of exogeneity of school choice is, for example, violated if students sort into schools based on anticipated shocks to school inputs (such as the change of a principal or changes in the school infrastructure); it is also violated if schools adjust their resources in response to the composition of the student body. Consequently, test score growth would be wrongly attributed to the school fixed effect rather than to student preparedness (see also the discussion by Mansfield, 2015). Model 1 includes a set of control variables at the student (classroom) as well as school level. The specification accounts for the most important observable dimensions of student sorting, in particular, sorting on ability (baseline test scores) and socio-economic status. The assumption is thus satisfied as long as the sorting does not occur on unobservable dimensions.[12] Once the strict exogeneity of school assignment is established, I assume that the student composition in terms of observable characteristics

[12] A detailed discussion on potential violations of such an exogenous mobility assumption is also provided by Kramarz et al. (2015).

is also pre-determined, so that Equation 14 holds.

3.3 Estimation

To estimate the model, I transform Equation 1 into a classroom-level equation. This transformation, which reduces the computational burden, is appropriate for two reasons. First, the variation of interest—i.e., the variation in school and teacher quality—occurs at the classroom level since each student has exactly one classroom teacher in a given year. Second, student characteristics enter into the model in a linear and additively separable way, such that they can be aggregated to the classroom level without loss of information.[13] Estimating the model at the classroom level instead of the student level reduces the dimension of the matrix that needs to be inverted in the estimation by a factor of about 20 (the average class size is 21). Thus, I estimate the following classroom-level model:

$$\overline{y}_{cjst} = \alpha_s + \mu_j + \overline{X}'_{ct}\gamma + V'_{jt}\delta + W'_{st}\rho + D'_t\psi + G'_t\xi + \epsilon_{cjst}, \tag{16}$$

where c denotes the classroom. The outcomes under study $\overline{y}_{cjst}$ are the average math and reading test scores of students in classroom c, which is taught by teacher j in school s at time t. The specification includes student control variables at the classroom level (average baseline test scores in math and reading and their squares, the fraction of students with missing baseline test scores, fraction female, fraction of students eligible for free or reduced-price lunch, fraction with missing information on free or reduced-price lunch, fraction of white students, fraction with parents whose highest degree is a high school degree or less, fraction with missing information on parental education, average age, fraction of students with limited English proficiency, fraction of gifted students, fraction of learning-disadvantaged students), time-varying teacher characteristics (experience in bins of less than one year, 1-2 years, 3-5

[13] The difference between the estimation at the classroom level and at the individual level is the weighting scheme. If class sizes were unbalanced in the data, then the aggregation at the classroom level could deliver misleading conclusions on the actual impact of teacher assignment. North Carolina, however, has strict guidelines on class sizes (the maximum class size is 30). Moreover, I exclude classrooms with more than 30 and less than 10 students, and I run representative specifications at the student level, the results of which differ only slightly from the classroom level estimations (results not shown).

years, 6-11 years, 12 or more years), time-varying school characteristics (average baseline test scores, fraction of students eligible for free or reduced-price lunch, fraction of white students), as well as class size, year dummies, and grade dummies.

I estimate the model using OLS, restricting the estimation sample to the largest connected set of schools, which contains 96 percent of the schools and 99 percent of the teachers in the initial sample (see Table A.2).[14] I then decompose the variance based on the *estimated* fixed effects and coefficients (Card et al., 2013).

As Abowd et al. (1999) and Andrews et al. (2008) point out, a two-way fixed effects estimation that relies on worker (here: teacher) moves should take a potential "limited mobility bias" into account. If teachers rarely move, or if schools have few movers, the estimates of both teacher and school effects can be biased. Imagine, for example, a teacher who moves once and who moves across schools of identical quality. By chance, he draws a bad classroom in one school and a good classroom in the other. If the quality estimates for these two schools were based on this mover alone, the school effects for these two schools would be biased. As the number of movers per school grows, or alternatively, as the number of years that a teacher spends in each of his/her schools grows, the school effects should approach their true value.

Biases in school effects can also translate into biases in teacher effects. The teacher effect for the stayers is the test score increase net of school effects and control variables; therefore, the teacher effects for the stayers depend negatively on the school effect by construction. Consequently, a positive bias in a school effect leads to a negative bias in a teacher effect, and the other way around, which can generate an artificial negative correlation between teacher and school effects.

To address both problems, I follow the procedure suggested by Andrews et al. (2008). I compute the school fixed effects only for schools that have at least 10 teachers who move during the sample period. I denote the teachers who move at least once in the sample period as "movers" and the schools that have at least ten movers as "high turnover schools". In order to preserve the connected set, the schools with less than 10 teacher movers are collapsed into a large reference school for the purpose of the estimation (see Andrews et al., 2008). I

[14] In order to compute the connected set, I use the command *felsdvreg* in *STATA*.

present the decomposition results for two samples—for the sample of all teachers and for the sample of movers (see the sample description in Table A.2). A Monte Carlo experiment, which I describe in detail in the appendix (Section A.3.2), confirms that the restriction to high turnover schools and movers reduces potential biases in teacher value-added and school effects.

3.4 Variance decomposition

I decompose the variance in test scores across classrooms into its variance and covariance components based on Model 16. I consider three main categories: student preparedness (modeled as a background index based on lagged test scores and socio-demographic characteristics, $\overline{X}'_{ct}\gamma$), teacher quality (measured as teacher value-added, μ_j, and teacher experience effects $V'_{jt}\delta$), and school quality (measured as the school fixed effects, α_s, and the school composition effects, $W'_{st}\rho$). Considering all three categories results in three variances and three covariances. While the variances capture the main effects in the model, the covariances capture the sorting of teachers to students, teachers to schools, and students to schools. I also report the contribution of the remaining variance and covariance terms, i.e. the contribution of grade and time fixed effects as well as the error term, such that all variance contributions add up to 100 percent.

4 Data

Data sets and variables. I use administrative records for the universe of school children in North Carolina's elementary schools (grades 3-5) for the years 1997-2011. The data is provided under a restricted use agreement by the North Carolina Education Research Data Center at Duke University. Based on randomized identifiers, one can link information on students, teachers, and schools, and track them over time. In constructing the data set, I closely follow Jackson (2013) and Rothstein (2017).

As the outcome, I use students' end-of-grade test scores in math and reading based on state-wide standardized tests, and I standardize the test scores at the year-by-grade level. In addition to test scores, the data contain student background characteristics (age, ethnicity, parental education, eligibility for free or reduced-price lunch) and information on student preparedness (classified as gifted students, academically disadvantaged, or with limited English proficiency). I extract prior-year test scores from end-of-grade test files or, where missing, from the "masterbuild" files.[15] The masterbuild files also contain pretest scores for grade-3-students; these test scores were collected before the students entered the elementary school level.[16]

The end-of-grade files contain an identifier for the test proctor, which in most cases identifies the classroom teacher. In order to exclude those proctors who were not the classroom teachers, I follow the procedure as suggested by, for example, Rothstein (2017). This procedure uses the personnel file, which contains information on a teacher's grade level as well as whether the teacher taught a self-contained classroom in a given year. I exclude those exams supervised by a teacher who was not listed as teaching a self-contained classroom in grades 3-5 in the given year. Based on this restriction, about 75 percent of teachers supervised their own classrooms. I exclude the 25 percent of classrooms who had proctors that were different from their classroom teacher.

Using the teacher identifiers, I add teachers' observable characteristics based on salary files. These files contain information on teacher qualifications (degree, the school and state where the degree was obtained, whether the teacher has a license, whether the teacher is certified) as well as a teacher's experience.

Panel balance. The data contain about 2.6 million student-year observations, 127,000 teacher-year observations (34,000 teachers), and 1,440 schools over a period of 15 years. The sample is not completely balanced with respect to teachers and schools (see Figure A.1).

[15] These files contain information that the school reports to the North Carolina Department of Public Instruction.

[16] Grade 3 pretests are missing in 2006 for math, in 2008 for reading, and in 2009-2011 for both math and reading. I, therefore, exclude grade 3 students from the analysis in 2006 and 2008-2011. In addition, some of the background variables may be reported in the end-of-grade file in some years, and in the masterbuild file in other years. For detailed information, please contact the author.

In particular, teachers enter and exit the data set during the time period. Each teacher is observed in the sample for 3.7 years on average. About 30 percent of teachers are only observed in one year. The data do not contain any information on the reasons why teachers drop out or are missing in certain years.[17] By contrast, the panel of schools is rather balanced. On average, each school remains in the sample for about 11 years. About 40 percent of schools participate in the sample for the entire period.

The identification of both teacher and school fixed effects relies on the movement of teachers across schools, which is illustrated in Figure A.2. Intuitively, the more frequently a teacher moves, the more accurately can his fixed effects be recovered, independently of the school where he teaches. About 20 percent of teachers switch schools during the sample period, and about 80 percent of the movers switch schools only once. Conditional on moving, teachers move on average 1.19 times.

Samples. Throughout this paper, I use both student-level data sets and teacher-level data sets. Table A.2 presents an overview of the different samples. I use a student-year level data set to present summary statistics and a raw decomposition of the test score variance into within-classroom, across-classroom-within-school, and across-school components. To perform the estimation of Model 16, I collapse the student-year level data set to the teacher-year (i.e., classroom) level. In the teacher-year level data set, all student characteristics are classroom averages. For the variance decomposition, I further restrict the sample to the high turnover schools and; I also create a sample that contains only movers (i.e., teacher who move at least once during the sample period). Since I consider heterogeneity in sorting patterns across regions, I restrict the decomposition sample to those schools that have non-missing information on the type of region where they are located.

Summary statistics. Table A.3 summarizes the sample and the variables used in the analysis. The majority of the students come from backgrounds with low levels of parental education and low socioeconomic status. 11 percent of the students' parents are high school

[17] The reasons may rank from both professional reasons (e.g., changes to private schools, obtaining a degree), but teachers may also move to a different state or take a leave of absence.

dropouts, and 46 percent of parents obtained a high school degree but no further education. With 47 percent of students who are eligible for free or reduced-price lunch, the socio-economic status of the students is below the federal average during the sample period.[18]

5 Results

5.1 Descriptive evidence of variance contributions

Table 1 presents descriptive evidence on the relative contributions of within-classroom variation, between-classroom-within-school variation, and across-school variation to the overall variation of test scores. The largest share of the variation in test scores is explained by within-classroom variation (77 percent of the total variation in math test scores, and 81 percent of the total variation in reading test scores). Only 12 percent of the variation in math test scores and 10 percent of the variation in reading test scores can be attributed to variation across classrooms; a similar share can be ascribed to variation across schools—11 percent of the variation in math test scores and 10 percent of the variation in reading test scores.

Table 1: Descriptive evidence: Components of test score variances

	(1) Math	(2) Reading
Total test score variance	0.98	0.98
Contribution to total test score variance (in %)		
Between-school variance	11%	10%
Between-classroom-within-school variance	12%	10%
Within-classroom variance	77%	81%

Note: Decomposition of raw variances in math test scores and reading test scores. The sample contains 2,584,712 student-year observations in 122,146 classrooms and 1,361 schools. Each school has on average 90 classroom-year-observations, and the average class size is 21. Test scores are standardized at the year-by-grade level with a mean of 0 and a standard deviation of 1. Only schools with at least 10 teacher moves in the sample period and with non-missing information on the type of region where the school is located are included in the sample. In the sample, the average math test score is 0.03, and the average reading test score is 0.02.

[18]See https://nces.ed.gov/programs/digest/d12/tables/dt12_046.asp.

This paper focuses on explaining the test score variation across classrooms and schools, which accounts for 23 percent of the overall variation in math test scores and for 20 percent of the overall variation in reading test scores. Since value-added models such as Model 1 assume that teacher inputs enter linearly into the education production function, teacher sorting cannot affect the within-classroom variation of test scores if all students in a classroom have the same teacher.

This paper also considers heterogeneity in teacher and student sorting patterns across regions, both within and across schools. Table A.4 depicts regional differences in raw variances as a first descriptive step. I distinguish between five types of regions, using the classification by the NCERDC: large and mid-size cities, urban fringe, towns, and rural areas. The contribution of the between-school variance in math test scores to the total test score variance is highest in large cities (14 percent in both math and reading), and lowest in rural areas (8 percent in math an 7 percent in reading). By contrast, there are only small differences across geographic areas in the contribution of the across-classroom variance within the same school to the overall test score variance. The following analysis explores the role of teachers' and students' sorting patterns in explaining the differences in variance contributions across regions.

5.2 Estimation of the education production

Before conducting the variance decomposition, I verify that the education production function delivers sensible estimates of the teacher effects, the school effects, and the coefficients on the observable characteristics (Tables A.6 and A.5).

The magnitude of teacher value-added from the estimation is strikingly in line with the results that use alternative ways to compute value-added. For teacher value-added in math, I find a standard deviation of 0.191 (Table A.5), i.e. an increase in teacher quality by one standard deviation corresponds to an increase in average student test scores by 0.191 standard deviations. The result is almost identical to the result by Rothstein (2017), who finds a standard deviation of teacher value-added in math of 0.192 in the same sample, using a different model and estimation approach.[19] Teacher value-added is with a standard deviation of 0.138 substantially lower in reading (Table A.5), which is also in line with the literature.[20]

Teacher experience, by contrast, contributes little to student test score outcomes. Raising the part of teacher quality that is due to experience by one standard deviation corresponds to an increase in average test scores by 0.018 standard deviations in math and 0.010 standard deviations in reading (Table A.5). Teacher experience matters most at the margin of less than versus more than one year of experience (see Table A.6); only five percent of the teachers in the sample, however, are novice teachers, such that experience does not account for much of the overall teacher quality variation.

Similarly, the unobserved component of school quality ("school effects") proves more important than the observed component of school quality (average student test scores at the school level, student demographics, and socio-economic composition at the school level). Increasing the unobserved components of school quality by one standard deviation is associated with an increase in test scores by 0.123 standard deviations in math and 0.103 standard deviations in reading. Changes in the composition of the student body, which are net of the classroom composition, translate into smaller changes in test score outcomes. An improvement in the observed component of school quality by one standard deviation maps into an increase in test scores by 0.017 standard deviations in math and 0.029 standard deviations in reading (Table A.5).

[19] Rothstein (2017) uses the same method as Chetty et al. (2014a) and accounts for drift in teacher value-added, i.e. teacher value-added may change over time. Moreover, Rothstein (2017) and Chetty et al. (2014a) estimate the model at the student level and then average across students; by contrast, I first average across classrooms and then run the model to compute value-added. Chetty et al. (2014a) find a standard deviation of teacher value-added of 0.163 in math and 0.124 in reading in a different school district (see p. 15 of their paper). Rothstein (2017) adds a note of caution to these estimates and the estimates from his replication in the North Carolina data, arguing that the models are not robust to selection on student unobservables.

[20] Rothstein (2017) finds an even lower value of teacher value-added of 0.118 in reading.

5.3 Results of the variance decomposition

5.3.1 Main results

Table 2 presents the results of the variance decomposition for the sample of high turnover schools and movers, which is the sample that performs best in the Monte Carlo experiment (see Section A.3.2). The results for the complete sample of teachers are similar and are shown in the appendix (Table A.7). Table 2 distinguishes between three groups of variance contributions: the main effects (Panel I), the sorting effects (Panel II), and the remaining variances (Panel III), i.e. variances and covariances of grade and year controls as well as the variance of the error term. All three groups add up to 100 percent of the total variance. I further subdivide the contribution of sorting into the contribution of teacher sorting to schools and students (Panel II.A) and student sorting to schools (Panel II.B).

The largest contribution to the variation in test scores comes from student preparedness (here: the average level of student preparedness in the classroom). Overall, student preparedness explains 52 percent of the variation in math test scores across classrooms and 58 percent of the variation in reading test scores (Table 2, Panel I). This result emphasizes the persistence of test score outcomes since baseline test scores are the main measure of student preparedness.

Teacher quality is the second most important input in the education production function. It explains, in total, about 16 percent of the variation in math test scores and 9 percent of the variation in reading test scores. School quality explains a smaller amount of the variance compared to teacher quality with only 7 percent of the variance in math and 6 percent of the variance in reading. The school quality measure is computed net of classroom composition and teacher quality; therefore, only school-specific factors such as school facilities and school management enter into the school quality effect.

Student sorting to schools is the third most important factor in explaining the variation in student test scores (Table 2, Panel II.B). It contributes 11 percent to the variance in math test scores across classrooms and 14 percent to the variance in reading test scores across classrooms. Thus, the importance of student sorting ranks just after the importance

Table 2: Variance decomposition (Math and Reading): Main results

	(1)	(2)	(3)	(4)
	Math		Reading	
Test score average	0.017		0.000	
Test score variance across classrooms	0.238	100%	0.205	100%
(I) Contribution of variances (main effects)	*0.176*	*74.3%*	*0.150*	*73.2%*
Var(student preparedness)	0.124	52.0%	0.119	58.1%
Var(teacher quality)	0.037	15.6%	0.019	9.4%
Var(school quality)	0.016	6.6%	0.012	5.8%
(II) Contribution of covariances (sorting)	*0.041*	*17.1%*	*0.038*	*18.6%*
(II.A) Teacher sorting to schools and students	0.014	5.9%	0.010	4.9%
2Cov(teacher quality, student preparedness)	0.018	7.5%	0.013	6.3%
2Cov(teacher quality, school quality)	-0.004	-1.6%	-0.003	-1.3%
(II.B) Student sorting to schools				
2Cov(student preparedness, school quality)	0.027	11.2%	0.028	13.6%
(III) Remaining variance and covariance terms	0.021	8.7%	0.017	8.3%
Number of school effects	657			
Number of teacher effects	5,630			

Note: The table shows results of the variance decomposition based on Model 16. The dependent variables are end-of-grade test scores in grades 3-5 in math and reading. All specifications control for year and grade dummies. School fixed effects are obtained only for high turnover schools (i.e., schools with at least 10 movers). The remaining schools are pooled into the reference category. The results presented here are for the high turnover schools and the sample of movers (i.e., teachers who move at least once during the sample period).

of teacher quality in math, and even before the importance of teacher quality in reading. In other words, better-prepared students systematically reap the benefits of better schools.

Teacher sorting (Table 2, Panel II.A) explains, in total, 6 percent of the variation in math test scores across classrooms and 5 percent of the variation in reading test scores across classrooms. Teachers primarily sort on student preparedness, measured as the average level of student preparedness in a classroom. By contrast, teachers do not positively sort on school quality once the average level of student preparedness in a classroom is accounted for. The covariance between teacher and school quality is close to zero.

Teacher sorting to classrooms with higher levels of average student preparedness can come from sorting within or across schools. Table 3 decomposes the covariance between teacher quality and student preparedness into a within-school and an across-school component. In order to disentangle the two effects, I compute the average teacher quality for each school, as well as each teacher's deviation from the average teacher quality in the teacher's school. I then compute two covariances, (1) the covariance between the average teacher quality in the school and classroom quality—a measure of student and teacher sorting across schools—and (2) the covariance between a teacher's deviation from the school average and classroom quality—a measure of teacher sorting within schools. Sorting across schools is, overall, more important than sorting within schools (4.7 percent versus 2.7 percent in math and 3.7 percent versus 2.6 percent in reading).

5.3.2 Heterogeneity of sorting patterns across regions

Sorting of teachers to schools and classrooms as well as student sorting to schools may not be equally important across geographic areas. Residential segregation and, therefore, segregation in students' backgrounds might be more severe in large cities compared to rural regions. For similar reasons, teacher sorting patterns might differ between rural and urban areas as well. Finally, teacher and student assignments to classrooms can also differ between geographic locations since principals might have different constraints and preferences in different areas.

Before discussing the variance contributions in detail, this section acknowledges the dif-

Table 3: Teacher sorting on student quality across and within schools

	(1)	(2)	(3)	(4)
	Math		Reading	
Test score average	0.017		0.000	
Test score variance across classrooms	0.238	100%	0.205	100%
2Cov(teacher quality, student preparedness)	0.018	7.5%	0.013	6.3%
across-school	0.011	4.7%	0.008	3.7%
across-classroom-within-school	0.007	2.7%	0.005	2.6%
Number of school effects	657			
Number of teacher effects	5,630			

Note: Decomposition of the covariance of teacher quality $(\hat{\mu}_j + V'_{jt}\hat{\delta})$ and student quality (estimated as the classroom's predicted performance, $\overline{X}'_{ct}\hat{\gamma}$) into a between- and within-school component for math test scores (Panel I) and reading test scores (Panel II). The between-school covariance is computed as $2 * Cov((\overline{\hat{\mu}_j + V'_{jt}\hat{\delta}}), \overline{X}'_{ct}\hat{\gamma})$, and the within-school covariance is computed as $2 * Cov((\hat{\mu}_j + V'_{jt}\hat{\delta}) - (\overline{\hat{\mu}_j + V'_{jt}\hat{\delta}}), \overline{X}'_{ct}\hat{\gamma})$.

ferences in the outcome levels across regions, as displayed in Table 4 for math and Table A.8 for reading. Overall performance is lowest in large cities (0.09 standard deviations below the sample average in math and 0.12 standard deviations below the sample average in reading) and highest at the urban fringe (0.11 standard deviations above the sample average in math and 0.10 standard deviations above the sample average in reading). Thus, the performance difference between cities and the urban fringe amounts to almost 20 percent of a test score standard deviation, on average. Students in mid-size cites and towns perform below the sample average but not as low as the students in large cities, and students in rural areas perform slightly above the sample average but not as well as students in the suburban areas.

Table 4: Variance decomposition of math test scores: Heterogeneity across regions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Large city		Mid-size city		Urban fringe		Town		Rural	
Test score average	-0.087		-0.056		0.109		-0.027		0.083	
Test score variance across classrooms	0.329	100%	0.258	100%	0.213	100%	0.225	100%	0.176	100%
(I) Contribution of variances (main effects)	*0.231*	*70.3%*	*0.193*	*74.8%*	*0.157*	*73.6%*	*0.166*	*73.7%*	*0.134*	*76.2%*
Var(student preparedness)	0.167	50.7%	0.141	54.9%	0.104	48.9%	0.115	50.9%	0.085	48.6%
Var(teacher quality)	0.040	12.3%	0.039	15.2%	0.036	16.7%	0.037	16.5%	0.034	19.5%
Var(school quality)	0.024	7.3%	0.012	4.7%	0.017	8.0%	0.014	6.2%	0.014	8.1%
(II) Contribution of covariances (sorting)	*0.068*	*20.8%*	*0.045*	*17.4%*	*0.039*	*18.2%*	*0.041*	*18.2%*	*0.017*	*9.7%*
(II.A) Teacher sorting to schools and students	0.005	1.4%	0.020	7.6%	0.011	5.2%	0.021	9.5%	0.012	6.9%
2Cov(teacher quality, student preparedness)	0.009	2.7%	0.022	8.4%	0.015	7.1%	0.024	10.5%	0.018	10.1%
2Cov(teacher quality, school quality)	-0.004	-1.3%	-0.002	-0.8%	-0.004	-1.9%	-0.002	-1.0%	-0.006	-3.2%
(II.B) Student sorting to schools										
2Cov(student preparedness, school quality)	0.064	19.4%	0.025	9.8%	0.028	13.0%	0.019	8.6%	0.005	2.8%
(III) Remaining variance and covariance terms	0.029	8.9%	0.020	7.8%	0.017	8.2%	0.018	8.2%	0.025	14.1%
Number of school effects	63		197		162		89		146	
Number of teacher effects	798		2,099		1,972		970		1,830	

The table shows results for the variance decomposition based on panel regressions (Model 16), restricting the sample to high turnover schools and to teachers who move at least once during the sample period. The outcomes are math end-of-grade test scores. All specifications control for year and grade dummies. In these specifications, school fixed effects are obtained only for schools with at least 10 moves. The remaining schools are pooled into the reference category, and the results are presented for schools with at least 10 moves and for different types of regions based on the location of the school.

Tables 4 and A.8 further provide insights into differences in sorting patterns across regions. Student sorting accounts for a large portion of the across-classroom variance in test scores, but this result is mostly driven by large cities and the urban fringe (see Panel II.B). In large cities, student sorting to schools explains 19 percent of the test score differences in math and 21 percent of the test score differences in reading across classrooms; in rural areas, by contrast, student sorting to schools explains only 3 percent of the test score differences in math and 2 percent of the test score differences in reading.

The opposite pattern holds for teacher sorting to students both within and across schools (Table A.9). Both teacher sorting across schools and teacher sorting within schools are less important in large cities compared to rural areas. As one explanation for this pattern, schools in rural areas might compensate for a smaller degree of school choice by allowing for more imbalances in classroom and teacher assignments within schools.

In sum, large cities not only have the lowest average test scores but also the highest variance in test scores. The joint contribution of both teacher and student sorting is very similar across large and mid-size cities, suburban areas, and towns. Rural areas seem to be the most balanced in terms of the distribution of teacher and school inputs across students, suggesting that the low density of schools in rural areas limits students' ability to sort. Since the evidence in this analysis does not allow conclusions on the causal mechanisms behind teacher and student sorting, further research is necessary to address questions of causality.

5.4 Simulation of counter-factual assignments

An alternative way to quantify the impact of teacher sorting is to contrast the distribution under the current sorting as observed in the data with the outcome distributions under alternative, i.e. counter-factual, teacher assignments across classrooms. This section considers three different types of counter-factual teacher distributions:[21] an equitable distribution of teachers within schools, an equitable distribution of teachers within schools and school districts, and an equitable distribution of teachers within schools and within the whole state of

[21] These allocations are considered thought experiments in this paper and not implementable policies.

North Carolina. By construction, all counter-factual assignments leave the average outcome unchanged because teacher and classroom inputs enter as additively separable inputs into the education production function (Model 16). Therefore, the test score gains that some of the classrooms realize must translate into test score losses for other classrooms.

Table 5: Simulation of counter-factual teacher assignments

	(1)	(2)	(3)	(4)
	Math test score distribution		Reading test score distribution	
	pc75-pc25	pc90-pc10	pc75-pc25	pc90-pc10
(I) Original teacher allocation				
Original test score distribution	0.656	1.252	0.609	1.166
(II) Simulation 1: Random teacher allocation within schools				
Simulated test score distribution	0.603	1.156	0.579	1.110
Difference to original	-0.053	-0.096	-0.030	-0.055
Change relative to original	-8%	-8%	-5%	-5%
(III) Simulation 2: Random teacher allocation within districts				
Simulated test score distribution	0.575	1.103	0.562	1.079
Difference to original	-0.081	-0.150	-0.047	-0.086
Change relative to original	-12%	-12%	-8%	-7%
(IV) Simulation 3: Random within state				
Simulated test score distribution	0.568	1.089	0.554	1.067
Difference to original	-0.088	-0.163	-0.055	-0.099
Change relative to original	-13%	-13%	-9%	-8%

Note: This table shows simulation results for math and reading test scores. The outcomes are the average differences in test score outcomes between a classroom at the bottom and a classroom at top quartile of the performance distribution ("pc75-pc25"), or between a classroom at the bottom decile and a classroom at the top decile of the performance distribution ("pc90-pc10"). I simulate three different counter-factual teacher assignments: random allocation of teachers within schools (Panel II), random allocation of teachers within school districts (Panel III), and random allocation of teachers within the whole state of North Carolina (Panel IV). Panel I presents summary statistics of the original distribution. The simulations are based on 100 random teacher draws (without replacement), and the results are averaged across all random draws. For details on the procedure, see also Section 5.4.

Table 5 shows the results of the simulations. The table reports how the assignment schemes affect the performance gap between a classroom at the 75^{th} and 25^{th} percentile of the outcome distribution (interquartile range, column (1) for math and column (3) for reading) and the performance gap between a classroom at the 90^{th} and 10^{th} percentile of the outcome distribution (column (2) for math and column (4) for reading). With respect to the interquartile range, I find the following results: Random assignment within schools reduces the performance gap by 0.05 standard deviations in math and 0.03 standard deviations in reading. This is an improvement of 8 percent in math and 5 percent in reading, relative to the original performance gap. Allowing for random allocations within school districts corresponds to a reduction of the interquartile range by 12 percent in math and by 8 percent in reading (0.08 standard deviations in math, 0.05 standard deviations in reading), and allowing for random allocations within districts translates into a reduction of the interquartile range by 13 percent in math and 9 percent in reading (0.09 standard deviations in math, 0.06 standard deviations in reading). Analyzing the gap between the classrooms at the 90^{th} and the 10^{th} percentile of the outcome distribution, I find results of similar relative magnitudes.

In summary, simulating the random assignments provides a clear picture of the amount of the test score inequality that can be attributed to teacher sorting. The exercise confirms the importance of sorting both within and across schools.

6 Discussion

The following section discusses the method and findings of the present study in light of the literature.

Method. The main identifying assumption of the present paper—exogenous mobility of teachers across schools, conditional on a number of student, teacher, and school characteristics —parallels the identifying assumption in the most recent value-added literature. Many current studies identify teacher value-added based on teacher switching across schools and classrooms (Chetty et al., 2014a,b, Rothstein, 2017, Bacher-Hicks et al., 2014). The underlying assumption is that conditional on student, classroom, school, and observable teacher characteristics, the movements of teachers across and within schools are as good as random movements. Chetty et al. (2014a,b) and Bacher-Hicks et al. (2014) show that specifications that rely on teacher switching produce reliable estimates of teacher value-added, although some bias from "dynamic sorting" between teachers and students (i.e. sorting based on unobservable time-varying student characteristics) might remain (Rothstein, 2017).[22]

While the main interest of the studies by Chetty et al. (2014a) and Rothstein (2017) is to validate measures of teacher value-added, the interest of the present study is a comprehensive decomposition of test scores. This study, therefore, draws upon an AKM-type model that jointly determines the teacher and school fixed effects. To assess potential biases in teacher and school value-added, I conduct a series of Monte Carlo experiments (Andrews et al., 2008). The AKM-type model delivers similar results for the estimated standard deviations of teacher value-added compared to the results by Chetty et al. (2014a) and Rothstein (2017), which are based on a different method to recover value-added.[23]

Results. In line with the literature, I find that teacher sorting is substantial despite North Carolina's state-wide pay schedule that allows for only small variations in teacher pay across and within schools. Several studies conclude that teacher pay is only one—and probably not the most important—factor that explains sorting patterns across schools. Other characteristics of the workplace, such as student ability, play an important role in teachers' labor market behavior (Clotfelter et al., 2011, Boyd et al., 2013). This finding is not even particular to the teacher labor market. Evidence on other occupations shows that workers' remuneration is only one, and sometimes a minor, variable that explains worker switches across firms (Fox, 2010, Bonhomme et al., 2016).

[22] The estimation procedure outlined in detail by Chetty et al. (2014a) proceeds in several estimation steps: In a first step, the authors compute a teacher-level residual as the predicted classroom-level test score net of student characteristics and school fixed effects; in a second step, they compute teacher value-added in year t as the predicted value from a regression of the teacher-level residual in year t on the teacher-level residuals prior to year t. The second step accounts for changes in teacher value-added over time and the fact that prior-year residuals predict current teacher value-added only with error.

[23] The standard deviations of teacher value-added in math are 0.19 in the present study, 0.19 as estimated by Rothstein (2017) with the same data set, and 0.14 as estimated by Chetty et al. (2014a) with a different data set. The standard deviations for teacher value-added in reading are 0.14 in the present study, 0.12 as estimated by Rothstein (2017), and 0.12 as estimated by Chetty et al. (2014a). It is important to note, however, that the goal of the present paper is to provide an aggregate analysis of sorting. This model should, at this point, not be used to evaluate individual teachers since assessing the validity of teacher-level value-added is beyond the scope of this paper. See also the discussion by Horváth (2015) on measuring teacher value-added in the presence of sorting.

As the main result of the paper, I show that teacher, school, and student sorting contribute substantially to the across-classroom variation in test score outcomes. Sorting explains 17 percent of the across-classroom variance in math test scores, and teacher quality accounts for one-third of this contribution. This fraction seems large, compared to results from other settings. In a related study, Mansfield (2015) investigates teacher sorting within and across North Carolina's public high schools. He finds that teacher sorting contributes 3.1 percent to the test score gap between a student at the top and the bottom decile of student preparedness. By contrast, I find that teacher sorting accounts for 13 percent of the test score gap in math and for 8 percent of the test score gap in reading between a classroom at the bottom and one at the top decile of the test score distribution. The results from the two studies, however, do not necessarily contradict each other. First, as Mansfield (2015) himself points out, the role of teacher-student sorting might differ across grade levels. Second, substantial heterogeneity across courses might exist; in the current paper, teacher quality seems more important in math compared to reading. At the high school level, where students have even more diverse subjects, an aggregate estimate of sorting might mask heterogeneity across subjects. Third, Mansfield (2015) focuses on explaining differences across individual students rather than differences across classrooms; this is appealing in a setting where each student has a variety of classroom teachers. By contrast, in the present paper, I solely exploit across-classroom variation in teacher quality. Further research, however, is necessary to understand differences in sorting at different levels of schooling as well as the underlying dynamics that can explain why such differences exist.

7 Conclusion

This paper studies teacher and student sorting across and within schools as a source of inequality in student test score outcomes, using a data set on public elementary schools in North Carolina. The analysis relies on an extended version of the worker-firm sorting model that was initially proposed by Abowd et al. (1999). In total, assortative matching among teachers, schools, and students contributes 17 percent to the across-classroom variation in test scores. Investigating the channels of sorting, I find that one-third of this contribution is due to teacher sorting, both within and across schools. Furthermore, I document heterogeneity in sorting patterns across regions. In large cities, student sorting to schools, rather than teacher sorting, is one of the main drivers of education inequality.

From a methodological perspective, further research might consider a more flexible model of education production. In particular, a critical assumption of the AKM-type approach is the additive separability between school, teacher, and student inputs. The model can be extended to allow for match effects between students and teachers or between teachers and schools. Recent econometric frameworks study match effects in the AKM framework (Bonhomme et al., 2016) as well as reassignment policies in the presence of match effects (Graham et al., 2007, 2010, Graham, 2011, Graham et al., 2014, 2016). Adopting such frameworks to the current problem provides a promising avenue for future research.

References

Abowd, J. M., F. Kramarz, P. Lengermann, and S. Perez-Duarte (2004): "Are Good Workers Employed by Good Firms? A test of a simple assortative matching model for France and the United States," Manuscript.

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999): "High wage workers and high wage firms," *Econometrica*, 67, 251–333.

Andrews, M. J., L. Gill, T. Schank, and R. Upward (2008): "High wage workers and low wage firms: negative assortative matching or limited mobility bias?" *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171, 673–697.

Angrist, J. D., S. R. Cohodes, S. M. Dynarski, P. A. Pathak, and C. R. Walters (2016): "Stand and Deliver: Effects of Boston's Charter High Schools on College Preparation, Entry, and Choice," *Journal of Labor Economics*, 34, 275–318.

Bacher-Hicks, A., T. J. Kane, and D. O. Staiger (2014): "Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles," Working Paper 20657, National Bureau of Economic Research.

Bifulco, R. and H. F. Ladd (2007): "School choice, racial segregation, and test-score gaps: Evidence from North Carolina's charter school program," *Journal of Policy Analysis and Management*, 26, 31–56.

Bifulco, R., H. F. Ladd, and S. L. Ross (2009): "Public school choice and integration evidence from Durham, North Carolina," *Social Science Research*, 38, 71–85.

Bonhomme, S., T. Lamadon, and E. Manresa (2016): "A distributional framework for matched employer employee data," Manuscript.

Boyd, D., P. Grossman, H. Lankford, S. Loeb, and J. Wyckoff (2006): "How changes in entry requirements alter the teacher workforce and affect student achievement," *Education Finance and Policy*, 1, 176–216.

Boyd, D., H. Lankford, S. Loeb, and J. Wyckoff (2013): "Analyzing the determinants of the matching of public school teachers to jobs: Disentangling the preferences of teachers and employers," *Journal of Labor Economics*, 31, 83–117.

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

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a): "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 104, 2593–2632.

——— (2014b): "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 104, 2633–79.

Clotfelter, C. T., H. F. Ladd, and J. Vigdor (2005): "Who teaches whom? Race and the distribution of novice teachers," *Economics of Education Review*, 24, 377–392.

Clotfelter, C. T., H. F. Ladd, and J. L. Vigdor (2006): "Teacher-student matching and the assessment of teacher effectiveness," *Journal of Human Resources*, 41, 778–820.

——— (2007): "Teacher credentials and student achievement: Longitudinal analysis with student fixed effects," *Economics of Education Review*, 26, 673–682.

——— (2011): "Teacher mobility, school segregation, and pay-based policies to level the playing field," *Education*, 6, 399–438.

Deming, D. J., S. Billings, and J. Rockoff (2014a): "School Resegregation, Educational Attainment and Crime: Evidence from the End of Busing in Charlotte-Mecklenburg," *Quarterly Journal of Economics*, 129, 435–476.

Deming, D. J., J. Hastings, T. Kane, and D. Staiger (2014b): "School Choice, School Quality and Postsecondary Attainment," *American Economic Review*, 104, 991–1013.

Dieterle, S., C. M. Guarino, M. D. Reckase, and J. M. Wooldridge (2015): "How do principals assign students to teachers? Finding evidence in administrative data and the implications for value added," *Journal of Policy Analysis and Management*, 34, 32–58.

Dobbie, W. and J. Fryer, Roland G. (2013): "Getting beneath the Veil of Effective Schools: Evidence from New York City," *American Economic Journal: Applied Economics*, 5, 28–60.

Feng, L. (2010): "Hire today, gone tomorrow: New teacher classroom assignments and teacher mobility," *Education Finance and Policy*, 5, 278–316.

Feng, L., D. Figlio, and T. Sass (2018): "School accountability and teacher mobility," *Journal of Urban Economics*, 103, 1–17.

Fox, J. (2010): "Estimating the Employer Switching Costs and Wage Responses of Forward-Looking Engineers," *Journal of Labor Economics*, 28, 357–412.

Graham, B. S. (2011): "Econometric Methods for the Analysis of Assignment Problems in the Presence of Complementarity and Social Spillovers," in *Handbook of Social Economics*, ed. by J. Benhabib, A. Bisin, and M. O. Jackson, North-Holland, vol. 1, chap. 19, 965–1052.

Graham, B. S., G. W. Imbens, and G. Ridder (2007): "Redistributive effects of discretely valued inputs," Manuscript.

——— (2010): "Measuring the Effects of Segregation in the Presence of Social Spillovers: A Nonparametric Approach," Working Paper 16499, National Bureau of Economic Research.

——— (2014): "Complementarity and aggregate implications of assortative matching: A nonparametric analysis," *Quantitative Economics*, 5, 29–66.

——— (2016): "Identification and efficiency bounds for average match output under conditionally exogenous matching," Research Paper 16-12, USC-INET.

Hanushek, E. A. (1997): "Assessing the Effects of School Resources on Student Performance: An Update," *Educational Evaluation and Policy Analysis*, 19, 141–164.

Horváth, H. (2015): "Classroom Assignment Policies and Implications for Teacher Value-Added Estimation," Manuscript.

Jackson, C. K. (2009): "Student demographics, teacher sorting, and teacher quality: Evidence from the end of school desegregation," *Journal of Labor Economics*, 27, 213–256.

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

Kalogrides, D. and S. Loeb (2013): "Different Teachers, Different Peers: The Magnitude of Student Sorting Within Schools," *Educational Researcher*, 42, 304–316.

Kramarz, F., S. Machin, and A. Ouazad (2015): "Using compulsory mobility to identify school quality and peer effects," *Oxford Bulletin of Economics and Statistics*, 77, 566–587.

Lankford, H., S. Loeb, and J. Wyckoff (2002): "Teacher sorting and the plight of urban schools: A descriptive analysis," *Educational Evaluation and Policy Analysis*, 24, 37–62.

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

Mortensen, D. T. (1986): "Job search and labor market analysis," *Handbook of Labor Economics*, 2, 849–919.

Mortensen, D. T. and C. A. Pissarides (1999): "New developments in models of search in the labor market," *Handbook of Labor Economics*, 3, 2567–2627.

Owens, A., S. F. Reardon, and C. Jencks (2016): "Income Segregation between Schools and School Districts," Working Paper 16-04, CEPA.

Reardon, S. F., D. Kalogrides, and K. Shores (2016): "The Geography of Racial/Ethnic Test Score Gaps," Working Paper 16-10, CEPA.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005): "Teachers, schools, and academic achievement," *Econometrica*, 417–458.

Rockoff, J. E. (2004): "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review*, 94, 247–252.

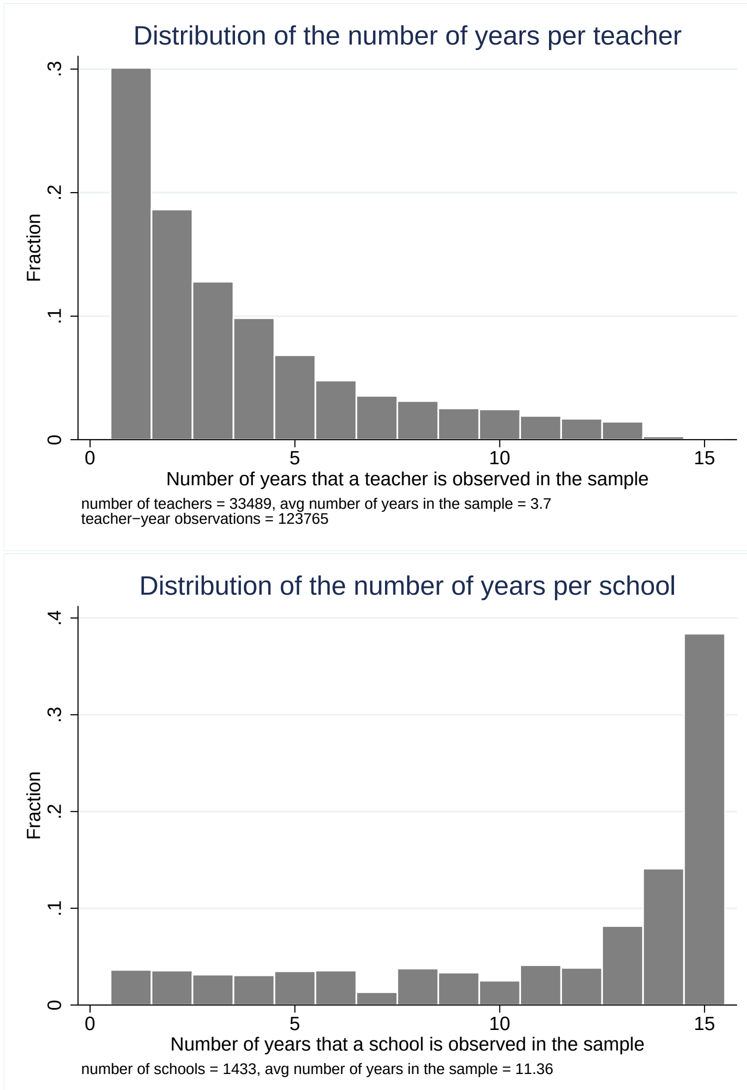
Rothstein, J. (2010): "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *The Quarterly Journal of Economics*, 125, 175–214.

——— (2017): "Measuring the Impacts of Teachers: Comment," *American Economic Review*, 107, 1656–84.

Sass, T. R., J. Hannaway, Z. Xu, D. N. Figlio, and L. Feng (2012): "Value added of teachers in high-poverty schools and lower poverty schools," *Journal of Urban Economics*, 72, 104–122.

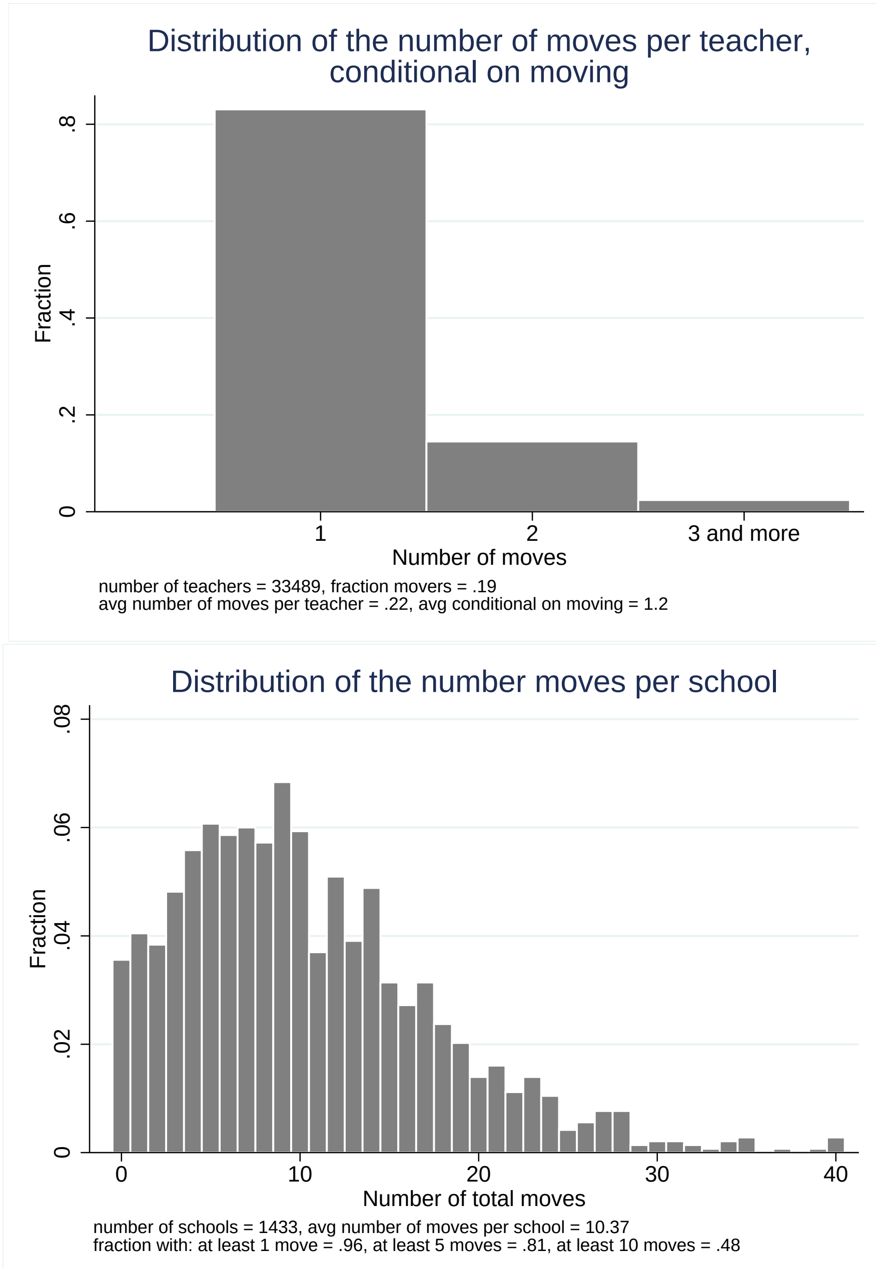
A.1 Figures

Figure A.1: Panel balance



Note: The figure displays the number of years that a teacher is present in the sample (top panel) as well as the number of years that a school is present in the sample (bottom panel).

Figure A.2: Moving frequencies



Note: The figure shows the number of moves per teacher, conditional on moving at least once (top panel) and the number of moves per school (bottom panel, inmoves and outmoves combined) for the estimation sample.

A.2 Tables

Table A.1: Variance components

(I) Variances (main effects)	
Var(student preparedness)	$\text{Var}(X'_{it}\gamma)$
Var(teacher quality)	$\text{Var}(\mu_{J(i,t)} + V'_{J(i,t)t}\delta)$
Var(school quality)	$\text{Var}(\alpha_{S(i,t)} + W'_{S(i,t)t}\rho)$
(II) Covariances (sorting)	
(II.A) Teacher sorting to schools and students	
2Cov(teacher quality, student preparedness)	$2\text{Cov}((\mu_{J(i,t)} + V'_{J(i,t)t}\delta), X'_{it}\gamma)$
2Cov(teacher quality, school quality)	$2\text{Cov}((\mu_{J(i,t)} + V'_{J(i,t)t}\delta), (\alpha_{S(i,t)} + W'_{S(i,t)t}\rho))$
(II.B) Student sorting to schools	
2Cov(student preparedness, school quality)	$2\text{Cov}(X'_{it}\gamma, (\alpha_{S(i,t)} + W'_{S(i,t)t}\rho))$

Note: This table provides an overview over the variance components of interest, as derived from a variance decomposition of Model 1. For details, see Section 3.

Table A.2: Samples

	(1)	(2)	(3)	(4)	(5)
Data set used for	Descriptives	Raw variances	Estimation	Decomposition	
				robustness	main
Level of observation	student-year	student-year	teacher-year	teacher-year	teacher-year
Sample restrictions					
Schools					
Type of region available	-	yes	-	yes	yes
High turnover schools	-	-	-	yes	yes
Largest connected set	-	-	yes	yes	yes
Teachers					
Only movers	-	-	-	-	yes
Number of observations					
Student-year	2,619,548	2,584,712	-	-	-
Teacher-year	123,765	122,146	122,896	77,349	26,614
Teachers	33,489	33,149	33,155	22,299	5,630
% of sample (1) teachers	100%	99%	99%	67%	17%
Schools	1,433	1,361	1,376	657	657
% of sample (1) schools	100%	95%	96%	46%	46%

Note: The table presents sample restrictions and resulting sample sizes for the samples used in this paper.

Table A.3: Summary statistics at the student-year level

	Mean	SD	Min	Max	Obs
Gender					
female	0.5	0.5	0	1	2,619,339
Age (in years)	10.37	1	7	17	2,619,548
Ethnicity					
white	0.6	0.49	0	1	2,619,548
black	0.28	0.45	0	1	2,619,548
hispanic	0.07	0.25	0	1	2,619,548
other	0.06	0.23	0	1	2,619,548
Parental education					
no high school	0.11	0.31	0	1	1,944,867
high school	0.47	0.5	0	1	1,944,867
up to community college	0.05	0.22	0	1	1,944,867
trade or business school	0.11	0.31	0	1	1,944,867
4-year college	0.21	0.41	0	1	1,944,867
graduate school	0.05	0.21	0	1	1,944,867
Free/reduced-price lunch					
eligible	0.46	0.5	0	1	2,352,214
Academically gifted					
gifted	0.13	0.33	0	1	2,613,106
Academically disadvantaged					
combined	0.06	0.23	0	1	2,615,731
reading	0.04	0.2	0	1	2,619,548
writing	0.04	0.19	0	1	2,619,548
math	0.02	0.14	0	1	2,619,548
Limited english proficiency					
yes	0.04	0.2	0	1	2,610,130
Baseline scores					
math	0.03	0.97	-17.44	4.47	2,080,597
reading	0.03	0.98	-17.24	5.24	2,111,377
math: missing	0.21	0.4	0	1	2,619,548
reading: missing	0.19	0.4	0	1	2,619,548
Outcomes					
reading score	0.02	0.99	-4.21	3.14	2,619,548
math score	0.03	0.99	-4.33	3.66	2,619,548
Teacher experience					
0 years	0.05	0.22	0	1	2,619,548
1-2 years	0.12	0.32	0	1	2,619,548
3-5 years	0.14	0.35	0	1	2,619,548
6-11 years	0.28	0.45	0	1	2,619,548
12 and more years	0.41	0.49	0	1	2,619,548
Class size	21.9	3.7	10	30	2,619,548
School composition					
Fraction FRL	0.41	0.24	0	1	2,619,548
Fraction ethnicity white	0.6	0.27	0	1	2,619,548
Average baseline test score reading	0.02	0.25	-1.24	1.65	2,619,548
Average baseline test score math	0.03	0.27	-1.62	1.74	2,619,548

Note: Information on free/reduced-price lunch is missing in 1997-1998. Information on parental education is missing in 2007-2011.

Table A.4: Descriptive evidence: Components of test score variances by type of region

	(1) Large city	(2) Mid-size city	(3) Urban fringe	(4) Town	(5) Rural
(I) Math test scores					
Test score average	-0.04	0.02	0.11	-0.03	0.03
Total test score variance	1.12	1.06	0.96	0.93	0.90
Contribution to total test score variance (in %)					
Between-school	14%	12%	10%	9%	8%
Between-classroom-within-school	15%	12%	12%	13%	12%
Within-classroom	71%	76%	79%	77%	80%
(II) Reading test scores					
Test score average	-0.07	0.01	0.09	-0.03	0.02
Total test score variance	1.09	1.05	0.96	0.94	0.92
Contribution to total test score variance (in %)					
Between-school	14%	12%	9%	8%	7%
Between-classroom-within-school	12%	10%	9%	10%	9%
Within-classroom	74%	78%	82%	81%	84%
Student-year-observations	203,966	653,633	663,470	386,934	676,709
Classroom-year-observations	9,682	30,906	30,297	19,058	32,203
Number of schools	83	307	281	249	441
Avg. # of classroom-year obs. per school	117	101	108	77	73
Avg. class size	21	21	22	20	21

Note: Decomposition of raw variances of math test scores and reading test scores into the between-school component, the between-classroom-within-school component, and the within-classroom component. The test scores are standardized at the year-by-grade level with a mean of 0 and a standard deviation of 1. The definition of the type of region is based on the geographic location of the school, and provided by the NCERDC. Only schools with at least 10 movers in the sample period and with non-missing information on the type of region are included in the sample. For schools that switch their type of region during the sample period, I use the oldest available type of region.

Table A.5: Standard deviations of estimated measures of student preparedness, teacher, and school quality

	(1)	(2)	(3)
	Measure	Standard deviation	
		Math	Reading
Student preparedness	$\overline{X}'_{ct}\hat{\gamma}$	0.352	0.345
Teacher quality			
Teacher value-added	$\hat{\mu}_j$	0.191	0.138
Teacher experience	$V'_{jt}\hat{\delta}$	0.018	0.010
School quality			
School effects	$\hat{\alpha}_s$	0.123	0.103
School composition	$W'_{st}\hat{\rho}$	0.017	0.029

Note: The table presents the standard deviations of the estimated measures of student preparedness, teacher quality (value-added and experience), and school quality (school effects, school composition in terms of student characteristics). The estimated measures come from an OLS estimation of Model 16. For definitions of the variables, see Section 3.3. For each of the estimated measures, I compute the standard deviation of the estimated measure. E.g., for teacher value-added ($\hat{\mu}_j$), I compute the standard deviation as $\sigma_{\hat{\mu}_j} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(\overline{\hat{\mu}}_j - \hat{\mu}_j)^2}$

.

Table A.6: Estimates of the education production function

	(1)	(2)	(3)	(4)
	Math		Reading	
	Coeff.	SE	Coeff.	SE
Baseline test score reading	0.232	(0.004)	0.362	(0.004)
Baseline test score math	0.367	(0.004)	0.250	(0.004)
Baseline test score reading (squared)	0.013	(0.006)	-0.013	(0.005)
Baseline test score math (squared)	-0.004	(0.005)	-0.010	(0.004)
Missing: Baseline test score reading	0.000	(0.003)	0.025	(0.003)
Missing: Baseline test score math	-0.008	(0.003)	-0.005	(0.003)
Female	-0.004	(0.008)	0.099	(0.007)
Eligible for free/reduced-price lunch (FRL)	-0.199	(0.008)	-0.205	(0.007)
Missing: FRL	-0.030	(0.008)	-0.024	(0.007)
Ethnicity white	0.228	(0.010)	0.219	(0.009)
Parental education high school or less	-0.057	(0.005)	-0.029	(0.005)
Missing: Parental education	0.016	(0.006)	-0.001	(0.005)
Age in years	-0.131	(0.006)	-0.109	(0.005)
Limited English proficiency	0.014	(0.014)	-0.286	(0.013)
Gifted student	0.530	(0.008)	0.419	(0.007)
Learning disadvantaged	-0.196	(0.013)	-0.485	(0.012)
Teacher experience (reference: 12 years or more)				
Teacher experience: 0 years	-0.086	(0.006)	-0.048	(0.005)
Teacher experience: 1-2 years	-0.020	(0.005)	-0.014	(0.004)
Teacher experience: 3-5 years	0.001	(0.004)	0.001	(0.004)
Teacher experience: 6-11 years	0.004	(0.003)	0.003	(0.003)
Class size	-0.009	(0.000)	-0.007	(0.000)
School: Fraction FRL	0.066	(0.013)	0.056	(0.011)
School: Fraction ethnicity white	0.067	(0.016)	0.119	(0.015)
School: Average baseline test score reading	-0.020	(0.010)	-0.049	(0.009)
School: Average baseline test score math	0.049	(0.009)	0.065	(0.008)
Grade (reference: grade 3)				
Grade 4	0.113	(0.007)	0.089	(0.006)
Grade 5	0.242	(0.012)	0.188	(0.011)
Teacher fixed effects	Yes			
School fixed effects	Yes			
Year dummies	Yes			
Number of classrooms	122,896			
Number of teacher fixed effects	33,155			
Number of school fixed effects	684			

Note: Parameter estimates based on Model 16. The dependent variables are end-of-grade test scores in math and reading. Test scores are standardized at the year-by-grade level. Analytic standard errors are in parentheses. The model includes both teacher and school fixed effects. The sample is restricted to the largest connected set of schools. School fixed effects are computed for schools with at least 10 in- or outmoves in the sample period. All remaining schools are pooled into the reference category. Therefore the number of school fixed effects is lower than the number of schools in the sample. Based on a sample of 1,376 schools and 33,155 teachers. Estimated using the command *felsdvreg* in *STATA*.

Table A.7: Variance decomposition (math and reading): All teachers (robustness)

	(1)	(2)	(3)	(4)
	Math		Reading	
Test score average	0.003		0.000	
Test score variance across classrooms	0.241	100%	0.210	100%
(I) Contribution of variances (main effects)	*0.186*	*76.9%*	*0.157*	*75.0%*
Var(student preparedness)	0.124	51.4%	0.120	57.1%
Var(teacher quality)	0.046	19.1%	0.026	12.4%
Var(school quality)	0.015	6.4%	0.012	5.5%
(II) Contribution of covariances (sorting)	*0.041*	*16.9%*	*0.040*	*19.1%*
(II.A) Teacher sorting to schools and students	0.016	6.6%	0.013	6.3%
2Cov(teacher quality, student preparedness)	0.022	8.9%	0.017	8.0%
2Cov(teacher quality, school quality)	-0.005	-2.3%	-0.003	-1.7%
(II.B) Student sorting to schools				
2Cov(student preparedness, school quality)	0.025	10.3%	0.027	12.8%
(III) Remaining variance and covariance terms	*0.015*	*6.2%*	*0.012*	*5.8%*
Number of school effects	657			
Number of teacher effects	22,299			

Note: The table shows results of the variance decomposition, based on Model 16. The dependent variables are end-of-grade test scores in grades 3-5 in math and reading. All specifications control for year and grade dummies. The decomposition sample contains only high turnover schools (i.e. schools with at least 10 teacher moves during the sample period).

Table A.8: Variance decomposition of reading test scores: Heterogeneity across regions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Large city		Mid-size city		Urban fringe		Town		Rural	
Test score average	-0.124		-0.070		0.095		-0.038		0.066	
Test score variance across classrooms	0.292	100%	0.235	100%	0.177	100%	0.187	100%	0.138	100%
(I) Contribution of variances (main effects)	*0.204*	*69.7%*	*0.167*	*71.1%*	*0.129*	*72.5%*	*0.140*	*74.9%*	*0.109*	*79.4%*
Var(student preparedness)	0.162	55.6%	0.136	57.7%	0.100	56.6%	0.109	58.6%	0.082	59.8%
Var(teacher quality)	0.022	7.6%	0.022	9.3%	0.017	9.5%	0.019	10.4%	0.017	12.2%
Var(school quality)	0.019	6.6%	0.010	4.1%	0.011	6.5%	0.011	5.9%	0.010	7.4%
(II) Contribution of covariances (sorting)	*0.070*	*24.1%*	*0.051*	*21.6%*	*0.034*	*19.1%*	*0.031*	*16.5%*	*0.007*	*5.0%*
(II.A) Teacher sorting to schools and students	0.010	3.3%	0.020	8.6%	0.007	4.1%	0.006	3.3%	0.004	2.7%
2Cov(teacher quality, student preparedness)	0.012	4.2%	0.021	8.8%	0.010	5.5%	0.010	5.4%	0.009	6.5%
2Cov(teacher quality, school quality)	-0.003	-0.9%	0.000	-0.2%	-0.003	-1.4%	-0.004	-2.0%	-0.005	-3.8%
(II.B) Student sorting to schools										
2Cov(student preparedness, school quality)	0.061	20.8%	0.031	13.0%	0.027	15.0%	0.025	13.2%	0.003	2.2%
(III) Remaining variance and covariance terms	*0.018*	*6.2%*	*0.017*	*7.3%*	*0.015*	*8.4%*	*0.016*	*8.6%*	*0.022*	*15.7%*
Number of school effects	63		197		162		89		146	
Number of teacher effects	798		2,099		1,972		970		1,830	

Note: The table shows results for the variance decomposition, based on panel regressions (Model 16). The outcomes are math end-of-grade test scores. All specifications control for year and grade dummies. The results are presented for schools with at least 10 moves during the sample period, and for teachers who move at least once during the sample period. The regional classification is based on the location of the school.

Table A.9: Teacher sorting on student quality across and within schools: Heterogeneity across regions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Large city		Mid-size city		Urban fringe		Town		Rural	
(I) Math test scores										
Test score average	-0.087		-0.056		0.109		-0.027		0.083	
Test score variance across classrooms	0.329	100%	0.258	100%	0.213	100%	0.225	100%	0.176	100%
2Cov(teacher quality, student preparedness)	0.009	2.7%	0.022	8.4%	0.015	7.1%	0.024	10.5%	0.018	10.1%
across-school	0.004	1.2%	0.013	4.9%	0.009	4.4%	0.015	6.5%	0.013	7.5%
across-classroom-within-school	0.005	1.4%	0.009	3.5%	0.006	2.7%	0.009	4.0%	0.005	2.6%
(II) Reading test scores										
Test score average	-0.124		-0.070		0.095		-0.038		0.066	
Test score variance across classrooms	0.292	100%	0.235	100%	0.177	100%	0.187	100%	0.138	100%
2Cov(teacher quality, student preparedness)	0.012	4.2%	0.021	8.8%	0.010	5.5%	0.010	5.4%	0.009	6.5%
across-school	0.006	1.9%	0.010	4.2%	0.006	3.3%	0.010	5.1%	0.006	4.6%
across-classroom-within-school	0.007	2.3%	0.011	4.6%	0.004	2.3%	0.000	0.2%	0.003	1.9%
Number of school effects	63		197		162		89		146	
Number of teacher effects	798		2,099		1,972		970		1,830	

Note: Decomposition of the covariance of teacher quality $(\hat{\mu}_j + V'_{jt}\hat{\delta})$ and student quality (estimated as the classroom's predicted performance, $\overline{X}'_{ct}\hat{\gamma}$) into an across- and within-school component, for math test scores (Panel I) and reading test scores (Panel II). The across-school covariance is computed as $2 * Cov(\overline{(\hat{\mu}_j + V'_{jt}\hat{\delta})}, \overline{X}'_{ct}\hat{\gamma})$, and the within-school variance is computed as $2 * Cov((\hat{\mu}_j + V'_{jt}\hat{\delta}) - \overline{(\hat{\mu}_j + V'_{jt}\hat{\delta})}, \overline{X}'_{ct}\hat{\gamma})$. The results are presented for different types of regions, based on the location of the school.

A.3 Technical appendix

A.3.1 Tests for teacher-school match effects

One identifying assumption of the AKM-type model is the additive separability of teacher and school fixed effects. To test whether this assumption is a reasonable approximation, I follow Card et al. (2013) and compare the model fit of Model 16 with the model fit of a model that includes the full set of teacher-by-school fixed effects. I estimate both models in a dataset that has at least two classrooms in each teacher-by-school cell, such that the teacher-by-school fixed effects have a meaningful interpretation. The resulting data set has 108,364 teacher-year (i.e. classroom) observations.

The model fit of the AKM-type model is almost as good as the model fit of the fully interacted model. The R^2 is 0.8449 in the AKM-type model and 0.8512 in the fully interacted model. The root mean squared error is 0.2008 in the AKM-type model and 0.2070 in the fully interacted model. I therefore conclude that the AKM-type model conveys a reasonable approximation and that teacher-school match effects are unlikely to affect the aggregate results on the contribution of sorting.

A.3.2 Monte Carlo experiment

I use a Monte Carlo experiment to assess whether the estimation recovers the true variances and covariances of teacher and school fixed effects, given the moving patterns in the data. As suggested by Abowd et al. (2004), I use an experiment that preserves the moving behavior that is observed in the data. I proceed as follows: First, I run Model 16 and save all coefficients on the observed characteristics. I use these coefficients to construct a fitted outcome net of school and teacher effects. Second, I remove the actual outcome (as observed in the data) from the data. Third, I randomly and independently draw teacher and school effects. The draws come from distributions which resemble the distributions of teacher and school effects as estimated in the first step (see the table notes of Table A.10 for details). Similarly, I draw idiosyncratic error terms (see the table notes of Table A.10 for details). Fourth, I construct a simulated outcome, based on the fitted outcome from the first step, the simulated teacher and school effects, and the simulated error terms. I run Model 16 again, but now substitute the actual outcome with the simulated outcome. Finally, I compute the variance-covariance matrix of the teacher and school effects from the regression outcome. I repeat the procedure 100 times, and for each type of school (more than one teacher move during the sample period, mover than five moves, more than 10 moves). Moreover, as suggested by Andrews et al. (2008), restricting the sample to movers (i.e. teachers who move at least once during the sample period) may reduce the bias in the person fixed effects. Therefore, I also compute the results separately for movers.

Table A.10 presents the results of the Monte Carlo experiment. Column (1) presents the median values of the variances in teacher and school effects as well as the median of their covariances based on the distribution of 100 “true” (i.e. simulated) teacher and school effects. Column (2) presents the medians from running Model 16 with the simulated outcome as the dependent variable. Column (3) shows the median of the differences in the variances and covariances from the 100 replications. Column (6) reports the mean relative bias. The three different panels show the results for the different schools samples (Panels I-III) as well as for movers only (Panel IV).

The sample restrictions help in correcting the variance estimates of both teacher and school effects towards their true variances. Considering the most restrictive sample of schools, i.e. schools with at least 10 teacher moves, reduces the mean relative bias in the variance of the school effect from 42 to 16 percent. Restricting the sample to movers reduces the mean relative bias in the variance of the teacher effect from 39 to 18 percent. The bias in the covariance between school and teacher effects almost disappears when using these restrictions. Thus, the sample restrictions suggested by Andrews et al. (2008) are a useful tool to reduce biases in the estimation of both unobserved school and teacher effects.

Table A.10: Monte Carlo experiment

	(1) Simulation median	(2) Monte Carlo median	(3) Median of difference	(4) Mean relative bias
(I) All schools				
Var(teacher effect)	0.040	0.055	0.016	39%
Var(school effect)	0.010	0.014	0.004	42%
2Cov(teacher effect, school effect)	0.000	-0.007	-0.007	-
(II) Schools with at least 5 moves				
Var(teacher effect)	0.040	0.054	0.014	35%
Var(school effect)	0.010	0.013	0.003	27%
2Cov(teacher effect, school effect)	0.000	-0.004	-0.004	-
(III) Schools with at least 10 moves				
Var(teacher effect)	0.040	0.053	0.013	33%
Var(school effect)	0.008	0.009	0.001	16%
2Cov(teacher effect, school effect)	0.000	-0.002	-0.002	-
(IV) Schools with at least 10 moves, only movers				
Var(teacher effect)	0.040	0.047	0.007	18%
Var(school effect)	0.009	0.010	0.002	17%
2Cov(teacher effect, school effect)	0.000	-0.002	-0.002	-

Note: The table shows results of a Monte Carlo experiment. Panel I presents school effects for all schools; panel II presents school effects for all schools with more than 5 teacher moves, and a joint school effect for the remaining schools with less than 5 moves; panel III computes school effects for all school with more than 10 moves, and a joint school effect for the remaining schools with less than 10 moves; panel IV presents the results for movers only. The data is is simulated as follows: I assume the teacher and school fixed effects to be normally and independently distributed with mean 0 and standard deviations $\sigma_j = 0.2$ for the teachers and $\sigma_s = 0.1$ for the schools. The error is defined as a composite error, which consists of shocks to teacher value-added, shocks to school value-added, and idiosyncratic shocks to classroom performance, i.e. $\epsilon_{ct} = \phi_{st} + \nu_{jt} + e_{cjst}$, where $\phi_{st} \sim N(0, 0.01)$, $\nu_{jt} \sim N(0, 0.02)$, and $e_{cjst} \sim N(0, 0.16)$. The simulation is based on 100 independent draws from these distributions. Column (1) shows the median of the simulated distribution in the data, column (2) shows the median of the distribution of effects as recovered based on Model 16 in the Monte Carlo experiment, column (3) shows the median of the distribution of differences between the estimate from the Monte Carlo experiment and the simulated effect, and column (4) shows the mean relative bias, where the mean is computed across all 100 draws.

A.3.3 Simulation of counter-factual teacher assignment schemes

To derive the outcome distribution under counter-factual teacher assignment schemes I implement the following simulation design. First, after estimating Model 16, I compute for every classroom the outcome net of teacher effects, i.e., I subtract the teacher effect from the observed outcome. Second, I create 100 random assignments of teachers to classrooms for three different counter-factual assignment schemes (i.e., random within schools, random within school districts, and random within the state of North Carolina). The teachers are drawn without replacements, and the random assignments are constructed separately for each year. Third, for each scenario, I compute 100 new test score outcomes for each classroom based on the observed outcome, net of the original teacher effect, and the 100 simulated teacher effects. I then average across all 100 draws. The results are reported in Table 5.