TEACHER QUALITY AND STUDENT INEQUALITY

Richard K. Mansfield

Address: Ives Hall, Room 266
Cornell University, Ithaca, NY 14853
E-mail: richard.mansfield@cornell.edu
Telephone: 781-724-1418 Fax: 607-255-4496

This paper uses eleven years of administrative data from North Carolina public high schools to examine the extent to which the allocation of teachers within and across public high schools is contributing to inequality in student test score performance. The existence of nearly 3,500 teacher transfers allows separate identification of each teacher's quality from other school-level factors. I find that teaching quality is fairly equitably distributed both within and across high schools: students among the bottom (top) decile of a student background index are taught by teachers who are, on average, at the 41st (57th) percentile of the value-added distribution.

I have greatly benefited from the input of Joseph Altonji, as well as Fabian Lange, Melissa Tartari, Costas Meghir, Lisa Kahn, Amanda Kowalski, and several anonymous referees. I also thank Mark Klee, Matthew Johnson, Priyanka Anand, Myrto Kalouptsidi, Rachel Heath, Gharad Bryan, Joseph Vavra, Chris Conlon, and Nicole Wright for valuable discussions, as well as seminar participants at Yale University. This research is based on data from the North Carolina Education Research Data Center at Duke University. I acknowledge the North Carolina Department of Public Instruction for collecting and providing this information.

1 INTRODUCTION

Recent research using matched student-teacher data has confirmed that teaching quality plays an important role in producing student test score improvement in elementary and middle schools (Rockoff [2004]; Hanushek *et al.* [2005]; Harris and Sass [2006]; Aaronson, Barrow, and Sander [2007]; Kane, Rockoff, and Staiger [2008]; Lockwood and McCaffrey [2009]). Furthermore, Chetty, Friedman, and Rockoff (2011) show that estimated teacher effects correspond to substantial impacts on adult earnings and other long-run outcomes. These findings have intensified concerns about the ability of underperforming schools to recruit and retain good teachers. One fears that the students who are already saddled with the least supportive parents, the most dangerous neighborhoods, and the most rundown schools will also be taught by the least effective teachers. However, the existing research has struggled to demonstrate convincingly the extent to which access to quality teaching is unequal.

This paper answers two questions: (1) How much does teaching quality vary across public high schools and across teachers within high schools? (2) To what extent are students who are otherwise disadvantaged more likely to attend the schools and the classes within schools with ineffective teachers?

The key challenge to characterizing inequality in the allocation of teacher quality has been isolating average teacher quality at a school from the effects of student sorting, principal quality, school facilities, and the surrounding neighborhood. Much of the early literature on teacher value-added focused only on within-school variation in teacher quality.[1] Other papers that do consider differences in teacher quality across schools generally either estimate an upper bound on the variance in teacher quality that also reflects differences in other school-level inputs,[2] or suffer from limited identifying variation (either small samples of schools or small samples of transferring teachers connecting each school).[3] In parallel work, Sass et al. (2010) do employ large samples of elementary-level teachers and schools in both North Carolina and Florida in an effort to determine whether high-poverty schools generally employ inferior teachers. However, they do not attempt to identify average teacher quality separately from other school-level inputs. They do test for the confounding influence of superior school-level inputs at low-poverty schools using differences in estimated effectiveness for teachers who move between high- and low-poverty schools. However,

1. See e.g. Rockoff (2004); Rivkin, Hanushek, and Kain (2005).
2. See e.g. Hanushek et al. (2005).
3. See e.g. Aaronson, Barrow and Sander (2007).

as demonstrated below, the validity of their test requires that an exogenous mobility condition be met, which they do not verify.[4] The only paper which provides a rigorous treatment of the conditions necessary for identification and precise estimation of school average teacher quality is parallel work by Kirabo Jackson (2013), whose focus is on teacher-school match effects rather than the impact of teacher allocation on student performance inequality.

One strand of the literature investigates the extent to which schools serving disadvantaged students hire teachers with inferior credentials.[5] Another examines whether such schools disproportionately lose their best teachers.[6] The analysis below combines the effects of both types of between-school teacher sorting along with within-school sorting into a comprehensive account of the contribution of the existing mechanism of teacher allocation to disparities between the performances of the least-prepared and best-prepared students. While other research has highlighted the potential bias in estimates of teacher quality that stems from non-random classroom assignment[7], this paper is the first to examine the extent to which students' classroom assignments contribute to the variation across students in average test performance over a high school career.[8] It also considers the impact of classroom assignments on the relative performance of disadvantaged students, to the extent that they are systematically assigned to classes with their schools' less effective teachers.

Furthermore, to this point nearly all of the attempts to examine the impact of teacher quality have used elementary or middle school test scores.[9] However, there are a number of advantages to using high school performance data to study the impact of teaching quality. First, we still have limited knowledge of how much the quality of teaching matters at the high school level. While a few studies have considered teacher quality in a high school context, they have either considered only Algebra 1 and English (e.g. Aaronson et al. [2007] and Jackson [2012]), or have focused only on the impact of particular observable teacher credentials (e.g. Clotfelter, Ladd and Vigdor [2007] and Xu, Hannaway, and Taylor [2011]). Second, teacher shortages tend to be far more severe at the high school level than at the elementary school level, and the subject-specific knowledge needed to be an effective teacher is greater. Thus, we have more reason to be concerned

4. Sass et al. (2010) also focus much of their effort on differences in the levels and marginal impacts of teacher experience and other observable teacher credentials across low- and high-poverty school environments.

5. See e.g. Lankford, Loeb, and Wyckoff (2002); Steele, Murnane, and Willet (2010).

6. See e.g. Hanushek et al. (2005); Boyd et al. (2007); Jackson (2009).

7. See e.g. Clotfelter, Ladd, and Vigdor (2006); Rothstein (2010).

8. Koedel (2010) does examine the collective effect of high school teachers on the probability of graduating high school.

9. These include: Hanushek et al. (2005); Boyd et al. (2007); Goldhaber, Gross, and Player (2007); Kane, Rockoff, and Staiger (2008); Kane and Staiger (2008); Lockwood and McCaffrey (2009); Jackson (2013); Rothstein (2010); and Sass et al. (2010).

about positive assortative matching that places inferior teachers in schools with the least-supported students. Third, high school teachers often teach four or five different classrooms each year, so that teachers may teach over 100 students per year. Hence, teacher impacts can be more precisely estimated.

I exploit administrative data from the North Carolina Education Research Data Center that permits high school students in the universe of North Carolina public high schools to be matched to their teachers and test scores in up to ten high school courses from 1997-2007. Such rich data permit identification and estimation of a flexible education production function that features both school-subject-specific intercepts and teacher-specific intercepts. Non-parametric identification of the joint distribution of teacher quality and school quality stems from a large network of nearly 3,500 teacher transfers, coupled with testable exogenous mobility and conditional random assignment assumptions. Such tests fail to reject the assumption of exogenous mobility, while tests for conditional non-random assignment of students to teachers find evidence of limited dynamic tracking, though not enough to introduce large biases into estimates. Given estimates of each teacher's quality, I then characterize the way quality teaching is currently being allocated. The next two paragraphs summarize my answers to the two questions posed above.

First, consistent with previous studies, I find considerable variation in teacher quality among North Carolina public high school teachers: a one standard deviation increase in teacher quality increases a student's expected test score by .18 student test score standard deviations, enough to move an average student from the 50th test score percentile to the 57th percentile. This estimate is generally consistent across subjects and fields (except for English I, where the estimated standard deviation in teacher quality is only about half as large). While 9% of the variation in student test scores is between schools, nearly all of this can be attributed to student sorting. Only about 1% of the total test score variation is potentially explainable by variation across schools in either school quality or average teacher quality. In fact, attending a school whose average teacher quality is one standard deviation better than the average school only increases expected test scores by .06 student test score standard deviations, holding other school-level inputs fixed. This is only enough to move an average student from the 50th test score percentile to the 52nd percentile. Moreover, variation in teacher experience across schools contributes almost nothing to across-school test score gaps. My analysis of the allocation of teachers to classes within schools indicates that most students tend to receive a mix of their schools' good and bad teachers across the courses they take, so that differences in quality among teachers from the same school only minimally contribute to

differences in average test score performance across students over their high school careers.

Second, I find that students whose observable background would predict low achievement do generally attend schools with lower average teacher quality, but that the magnitudes of these differences are fairly modest. Similarly, I find that such disadvantaged students are only slightly more likely to take classes with the relatively ineffective teachers at their schools. Overall, I find that students among the bottom (top) decile of a regression index of student background are taught by teachers who are, on average, at the 41st (57th) percentile of the value-added distribution. This gap, combined with smaller gaps in effective teacher experience, accounts for 3.0% of the high school performance gap between the top and bottom deciles.

The remainder of the paper is structured as follows. Section 2 presents the educational achievement production function I estimate. Section 3 discusses identification of the parameters of the function. Section 4 describes the data. Section 5 presents the estimation strategy. Section 6 presents the estimated distributions of teacher quality and school average teacher quality, both pooled across subjects and disaggregated by subject. Section 7 tests the conditional random assignment assumption and examines the model's out-of-sample predictive power. Section 8 examines how quality teaching is allocated within schools and quantifies its impact on the distribution of average test scores over students' high school careers. Section 9 examines the contribution to achievement inequality of the existing allocation of teachers to students. Section 10 interprets the findings and concludes.

2 THE EDUCATION PRODUCTION FUNCTION

Let Y_{ict} represent the standardized test score of student i in course c in year t. To minimize the impact of different choices of scales for exams taken in different subjects, Y_{ict} has been standardized relative to the appropriate course-year-specific state distribution. Let $s(i,t)$ represent the school that student i attended in year t, and let $r(i,c,t)$ represent the teacher who taught student i in course c at time t. I estimate the following specification of the production function for standardized achievement:

$$Y_{ict} = \tilde{\mathbf{Y}}_{\mathbf{i}}^{\mathbf{t-1}}\alpha_{\mathbf{c}} + \mathbf{X_{ict}}\beta_{\mathbf{c}} + \delta_{s(i,c,t)c} + d(ex_{r(i,c,t)t}) + \mu_{r(i,c,t)} + \epsilon_{ict} \tag{1}$$

The parameters of interest in this specification are those relevant to evaluating the contribution of high school teacher inputs to student inequality: the set of persistent teacher qualities $\{\mu_r\}$ and the profile of teacher growth with experience $d(*)$.

In recognition of research indicating considerable persistent unobserved heterogeneity in teachers' performance, each teacher is assumed to have his/her own baseline ability to impact test scores, captured by $\mu_{r(i,c,t)}$. Note that a teacher's ability to increase test scores is assumed to be common across courses. This restriction is necessary for pooling the information about school quality obtained from teachers of different subjects who transfer between schools. It will be discussed in further detail in the next section.[10] To the extent that a teacher's quality is subject-specific (i.e. the teacher is better at teaching Algebra 1 than Algebra 2), we can interpret our estimates of a teacher's quality as capturing the teacher's average ability to increase test scores across the subjects taught during the sample period, weighted by the frequency with which the teacher actually taught those subjects. $\mu_{r(i,c,t)}$ is treated as a parameter to be estimated (one for each teacher), and I will generally refer to $\mu_{r(i,c,t)}$ as teacher quality.

The function $d(ex_{r(i,c,t)t})$ captures predictable growth in teacher effectiveness with experience, $ex_{r(i,c,t)t}$. The function is assumed to be common across teachers and courses. In practice, $d(ex_{r(i,c,t)t})$ will be flexibly parametrized using indicators for narrow ranges of experience.

$\delta_{s(i,t)c}$ captures the collective impact of all persistent inputs or conditions that are specific to the school-course combination that affect student learning independently of teacher quality. It reflects school-wide factors such as principal quality, safety of the neighborhood, and the quality of school facilities, as well as school-course-specific inputs like the quality of the textbook the school uses to teach a particular course. $\delta_{s(i,t)c}$ will be estimated using a full set of school-course fixed effects. As discussed in Section 3, $\delta_{s(i,t)c}$ may also reflect non-random sorting of students into schools and courses driven by unobserved student characteristics.

This specification for the contribution of school and teacher inputs to achievement offers a couple of desirable features. First, the joint distribution of unobservable persistent teacher quality (μ) and unobservable persistent school-course effects (δ) is left unrestricted, allowing for arbitrary sorting of teachers into schools and courses within schools.

Second, persistent teacher skill is captured by a single dimension, which enables a tractable discussion of the distribution of teacher quality across students. However, the ability of a teacher

10. Note that certification in North Carolina is specific to the field (i.e. math or science) and not to the course (i.e. geometry or chemistry), suggesting that the state believes teachers can teach different courses within a field interchangeably.

to raise test scores is nonetheless allowed to vary over time as experience accumulates.

$\tilde{\mathbf{Y}}_{\mathbf{i}}^{\mathbf{t-1}}$ is a vector of English and math test scores from 7th and 8th grade and their squares. This vector of prior test scores is intended to serve as a proxy for the impact of prior inputs on student test scores. Similarly, I use a vector of observable student, family, and classroom characteristics, denoted $\mathbf{X_{ict}}$, as a proxy for current student, family, and peer inputs. Other research using achievement in multiple subjects has exploited cross-course student fixed effects to control for student selection into classrooms (e.g. Clotfelter, Ladd, and Vigdor (2010) or Xu, Hannaway, and Taylor (2011)). The use of both teacher fixed effects and school-course fixed effects rendered this approach computationally infeasible. Note, though, that the coefficients on both past test scores and student and classroom characteristics are course-specific. This allows the weight given to an 8th grade math score to be greater for a high school math course than an English course, thus allowing the pattern of past test scores to reveal comparative advantages of particular students for particular subjects.[11] Neither $\{\beta_c\}$ nor $\{\alpha_c\}$ represent parameters of interest in this specification. They are included as a control function to partially purge the remaining parameters of the model of the impact of student sorting among teachers within a school. The impact of using proxies for the contributions of current and past student, family, and peer inputs is discussed further in the next section.

The error term, ϵ_{ict}, is assumed to be composed of the sum of three components:

$$\epsilon_{ict} = \phi_{s(i,t)t} + \nu_{r(i,c,t)t} + e_{ict} \tag{2}$$

The first error component, $\phi_{s(i,t)t}$, represents the transient component of school inputs. It captures fluctuations in principal quality, crime waves, renovations of school facilities, etc. Idiosyncratic shocks to school quality are assumed to be common to all courses within the school.

The second error component, $\nu_{r(i,c,t)t}$, captures idiosyncratic year-specific deviations in a teacher's performance from the path defined by his/her baseline ability and experience. Such deviations might be caused by fluctuations in teacher health, personal obligations, or even the extent to which the standardized test in a given year happens to focus on the content the teacher covers most effectively or intensively.

11. One possible concern is that the control set of demographics and past math and reading test scores may do a poor job in predicting student preparation and talent in a particular subject. In practice, the standard deviation in the regression index of observables $X_{ict}\beta_c + \tilde{Y}_i^{t-1}\alpha_c$ is between .6 and .78 of a test score standard deviation across all 10 subjects. Overall, past test scores account for 57% of the variation in high school achievement, which suggests that such scores are capturing the bulk of persistent student ability.

Finally, e_{ict} captures both (1) the component of the combined impact of past and current student, parent, and peer inputs that could not be predicted given the observable proxies, $\tilde{Y}_i^{t-1}$ and X_{ict}, and (2) idiosyncratic student-course-specific measurement error in test scores (i.e., the extent to which the student's performance on the day of the exam fails to capture the student's underlying mastery of the content the exam is designed to test).

Standard errors are adjusted for the existence of both $\phi_{s(i,t)t}$ and $\nu_{r(i,c,t)t}$, and tests of conditional random assignment and endogenous teacher mobility presented in Section 6 and Section A5 of the Web Appendix will address potential biases that could result from failure to explicitly model how teacher assignments respond to $\phi_{s(i,t)t}$, $\nu_{r(i,c,t)t}$ and e_{ict}.

3 IDENTIFICATION

As emphasized by Todd and Wolpin (2003) and Meghir and Rivkin (2010), among others, endogenous choice of inputs by students, parents, and schools represents a formidable obstacle to identifying the parameters of an education production function. The first potential issue is the parents' endogenous choice of school. In the likely scenario that there is some degree of positive assortative matching between unobservably superior students and superior schools, the school-subject fixed effect parameters, $\hat{\delta}_{sc}$, will capture a combination of true school quality and student quality, and will have no causal interpretation. Similarly, students' endogenous choice of courses (for those courses which are optional) implies that variation in $\hat{\delta}_{sc}$ among courses within a school will likely reflect both true subject-specific variation in school quality (e.g. textbook choice) and variation in average (unobserved) student quality among subjects. Importantly, however, since μ_r is identified purely from comparisons among teachers while they are teaching in the same course at the same school, student selection into schools and courses will not bias estimates of teacher quality and the teacher experience profile. The school-course fixed effects thus effectively act as a flexible control function that eliminates bias from any persistent patterns of student sorting into schools and courses.

The second potential endogeneity problem is that students may not be randomly assigned to teachers within schools, so that the average test scores of a given teacher's students may partly reflect deviations in student inputs from school averages. Rothstein (2010), in particular, has argued that this problem is severe at the elementary school level. In order to recover unbiased estimates of persistent teacher quality, μ_r, we need the identity of a student's teacher to provide no further infor-

mation about the student's unobservable current or prior inputs, given the information contained in observable prior test scores, observable current inputs, and the school-course combination in which the student is enrolled:

Assumption 1: Conditional Mean Independence of Students' Unobserved Inputs and Teacher Identity

$$E[\epsilon_{ict}|r(i,c,t) = r', (s(i,t),c) = (s',c'), \tilde{\mathbf{Y}}_i^{t-1}, \mathbf{X}_{ict}] =$$

$$E[\epsilon_{ict}|(s(i,t),c) = (s',c'), \tilde{\mathbf{Y}}_i^{t-1}, \mathbf{X}_{ict}] \; \forall \; r' \in \mathcal{R}, (s',c') \in \mathcal{SC}, \tag{3}$$

where ϵ_{ict} is the composite error term defined in (2), $\mathcal{R}$ is the set of all teachers, and $\mathcal{SC}$ is the set of all school-course combinations.

At the high school level, constructing students' schedules is an onerous task performed by schedule-making computer programs, making it difficult for principals to assign individual students to teachers, and for students to target particular teachers. However, principals must still assign teachers to difficulty levels, and students may be choosing whether to take an honors class based on private information about deviations in their own expected inputs from past levels. Thus, the validity of Assumption 1 depends critically on the extent to which the sorting of students into levels is captured by prior test scores and observable current inputs. Indeed, research by Jackson (2012) shows that failure to control for track may generate non-trivial bias in teacher value-added estimates.

While the full set of observed inputs is presented when I discuss the data in Section 4, it is worth noting here that I include in $\mathbf{X}_{ict}$ the average prior test scores and average demographic characteristics of the other students in student i's class and an indicator for whether the class is designated as honors/AP. These measures are likely to serve as effective proxies for the difficulty level of a class, thereby mitigating bias stemming from violation of conditional random assignment of students to teachers.[12] Nonetheless, because dynamic sorting of students to teachers has been such a major concern in the literature, I also test Assumption 1 directly in Section 7 using the procedures introduced by Chetty et al. (2011) and Jackson (2012). The results of these tests do suggest some limited dynamic sorting even after controlling for classroom characteristics and track; however, as

12. Note that the classroom characteristics and track indicators are allowed to have subject-specific coefficients, which permits heterogeneity in sorting on student unobservables across subject-difficulty level combinations.

I describe below, dynamic sorting of this magnitude is unlikely to be an important source of bias for characterizing the contribution of teacher allocation to student achievement inequality.

A third potential endogeneity problem is also ruled out by Assumption 1. Even if students are conditionally randomly assigned to teachers, students (and parents) may respond to the quality of teaching they receive by adjusting their own current inputs. For example, a student saddled with an ineffective teacher may be more likely to study the textbook harder or pay for tutoring services. Indeed, Pop-Eleches and Urquiola (2011) document that parental effort toward student achievement in Romania declines when a student obtains access to a better school environment. Such input compensation would cause estimates of teacher quality to be muted in magnitude, and the variance in teacher quality to be underestimated.[13] However, we may have less reason to be concerned that parental inputs compensate for teacher quality at the high school level than at the elementary and middle school levels. For example, if a first grade teacher fails to teach a child to read, the child's parents can probably teach the child to read at home. In contrast, most parents are likely to be far less comfortable explaining physics concepts. In this case, their only option would be to pay for costly professional tutoring. Furthermore, since such input compensation is caused by the teacher's quality, for some questions the parameter of interest is the raw teacher contribution combined with the average compensation component of e_{ict} for the teacher.[14] We could formalize this by decomposing $e_{ict} = e_{1ict} + e_{2ict}$, where e_{1ict} represents the component of student/parent inputs that would be delivered regardless of the teacher assigned, and e_{2ict} represents the component of student/parent inputs that is contingent on the identity of the student's teacher. Then we could weaken Assumption 2 above by replacing ϵ_{ict} with ϵ_{1ict}, where ϵ_{1ict} contains all the components of ϵ_{ict} except e_{2ict}, and reinterpret the estimated teacher effects as capturing $\tilde{\mu}_{r'} = \mu_{r'} + E[e_{2ict}|r(i,c,t) = r']$.

A fourth potential endogeneity problem stems from the possibility that teachers choose the effort they make or the content they teach. I do not explicitly model teacher effort. Instead, I

13. The assumption of additive separability of teacher inputs from student inputs implies that the two are substitutes. However, if in fact teacher inputs and student inputs are complements, students might increase their inputs in response to a particularly effective teacher, and the bias would be reversed.

14. For example, if we were interested in predicting what would happen to the relative achievement of two classrooms of students if we swapped their teachers, then the estimates that include input compensation might be appropriate. If, on the other hand, we were interested in predicting the performance of the same two teachers in an alternative school environment where students had very little time at home, then we might prefer estimates that exclude input compensation, since the degree of input compensation among a teacher's students is less likely to remain stable across very different contexts. Previous versions of this paper featured specifications with parameters capturing school-average sensitivity to teacher quality that made estimated teacher effects robust to variation in input compensation across schools, and yielded very similar results for the contribution of teacher allocation to student performance inequality.

assume that teachers do not systematically adjust effort in response to school, student, or peer inputs. With this assumption, persistent differences in effort across teachers will simply represent an important component of persistent quality μ_r, and idiosyncratic deviations in effort from a teacher's norm will be captured by ν_{rt}. A related concern is that persistent differences in teacher performance may be reflecting the extent to which teachers adhere to the state curriculum rather than differences in ability to foster learning. Fortunately, several aspects of the context surrounding the data help allay these fears.[15]

Notice that since the school-year idiosyncratic deviation $\phi_{s(i,t)t}$ is a component of $\epsilon_{i,c,t}$, Assumption 1 also rules out the possibility that certain teachers teach a disproportionate fraction of their career during the period in which their school is experiencing its relatively ineffective years compared to its full-sample average quality. Clearly, some teachers will happen to teach more in their schools' down years, so that their relative quality will be underestimated compared to the school average. If disadvantaged subpopulations of students are disproportionately present during their schools' less effective years, I could underestimate the quality of the teachers they receive. Consequently, in Section 8 I explicitly account for the set of teachers that were actually present when examining whether such subpopulations are systematically being assigned their schools' relatively ineffective teachers.

Note, however, that the average value of ϕ_{st} experienced among all teachers ever teaching at the school in the sample is zero by construction, since the δ_{sc} parameters will capture the mean quality of the school (and course). Thus, if the set of teachers who transfer across schools is not systematically related to year-specific deviations in average quality of the schools they leave or arrive at, then if there are enough transferrers connecting a given school to the network, the average bias among transferring teachers will tend to zero, and the estimate of the average quality of the school's teachers will be unbiased.

Indeed, the same point can be made with respect to non-random sorting of students to teachers within schools. The average bias in $\hat{\mu}_r$ among all teachers teaching the same course at the same school must be zero by construction, since school-course averages of unobserved student-level inputs will be captured by the school-course-specific intercept, δ_{sc}, and every student must have been

15. First, in recent years No Child Left Behind legislation has put pressure on principals to ensure that teachers teach the standard curriculum, since schools that fail to meet state standards are subject to sanctions and possible closure. Second, the North Carolina end-of-course exam scores I use as outcome measures must comprise 25% of the student's year-end grade in a given subject, so that parents are likely to complain about teachers that ignore the standard curriculum. Finally, during the sample period in North Carolina, teacher bonuses of up to $1,500 were linked to average test scores of the students in the school at which they teach. Thus, teachers are under considerable pressure to teach the tested material.

taught by *some* teacher. Consequently, if transfer decisions are unrelated to the *bias* in transferring teachers' quality estimates,[16] violations of Assumption 1 driven by non-random student sorting within schools will not bias estimates of average teacher quality.

More generally, if we are only interested in consistently estimating average teacher quality at each school (or, more specifically, each school-course combination), we can replace Assumption 1 with a weaker exogenous mobility condition.

Assumption 2: Exogenous Mobility

$$
\begin{aligned}
&E[\epsilon_{ict}|r(i,c,t) \in \tilde{\mathcal{R}}_{(s',c')}, (s(i,t),c) = (s',c'), \tilde{\mathbf{Y}}_i^{t-1}, \mathbf{X}_{ict}] \\
&= E[\epsilon_{ict}|(s(i,t),c) = (s',c'), \tilde{\mathbf{Y}}_i^{t-1}, \mathbf{X}_{ict}] \; \forall \; (s',c') \in \mathcal{SC} \\
&E[\epsilon_{ict}|r(i,c,t) \in \tilde{\mathcal{R}}_{(s',c')}, (s(i,t),c) \in \mathcal{SC} \setminus (s',c'), \tilde{\mathbf{Y}}_i^{t-1}, \mathbf{X}_{ict}] \\
&= E[\epsilon_{ict}|(s(i,t),c) \in \mathcal{SC} \setminus (s',c'), \tilde{\mathbf{Y}}_i^{t-1}, \mathbf{X}_{ict}] \; \forall \; (s',c') \in \mathcal{SC}, \qquad (4)
\end{aligned}
$$

where $\mathcal{SC}$ denotes the set of all school-course combinations, and $\tilde{\mathcal{R}}_{(s,c)}$ denotes the set of teachers observed transferring either into or out of school-course combination (s,c). The first condition in Assumption 2 states that teachers who transfer into or out of any particular school-course combination (s',c') on average do not teach students with above or below average unobserved inputs while at (s',c'), conditional on the identity of the school, the course, and the level of observable student inputs. The second condition states that the same set of transferring teachers are also not disproportionately assigned to students with above or below average unobserved inputs while teaching at the alternative school-course combination they taught at before or after (s',c'), conditional on the identity of the alternative school and course, and the level of observable student inputs.

The intuition behind the need for exogenous mobility is as follows. A teacher's persistent quality, μ_r, can be identified relative to the other teachers teaching her course at her school by comparing the average residuals of her students' test scores with those of the other teachers, after removing the predicted impact of student- and classroom-level inputs.[17] If a teacher has taught at multiple schools, then she can be placed in the distribution of within-school teacher quality in both schools. Likewise, a teacher who has taught multiple subjects within a school can be placed in the

16. The two would be related, for example, if teachers who consistently received unobservably bad students at a school became disgruntled and thus more inclined to transfer.

17. This assumes that all teachers have taught a large number of students, so that sampling error from measurement error and average levels of unobserved student-level inputs is minimal.

distribution of within-school teacher quality in both subjects. With only one teacher linking schools (subjects), we would need to assume that her ability to increase test scores is the same across the two schools (subjects) in order to infer relative school (subject) quality from her students' relative performance at the two schools (subjects). Once relative school and subject qualities are known, we can shift the within-school distributions of average student residuals appropriately to place the performance of teachers from all school-subject combinations in a neutral teaching environment. However, with many linking teachers, relative school quality is identified by differences in the *average* performance of transferring teachers across the two schools. Similarly, a school's comparative advantange in a particular subject is identified by differences in the *average* performance of transferring teachers across the two subjects. Thus, Assumption 2 only requires that transferring teachers are not *systematically* more productive at one of the schools (or subjects), so that the difference in the average performance of students taught by transferring teachers still identifies relative school (or subject) quality.

Note that several obvious forms of systematic teacher mobility do not constitute violations of Assumption 2. These include systematic transfer of teachers to better schools (or to a school's relatively good subjects), increased probability of transfer among inferior or superior teachers, or a combination of relatively good teachers being disproportionately likely to move to good schools. This is because the average performance among students taught by transferring teachers at each school and subject under such mechanisms is predictable given knowledge of the teacher and school-subject combination. Knowing that the teacher who teaches a given student transferred to the school or the subject does not provide any further information about any component of the composite error in these contexts.

There are, however, several mechanisms by which the exogenous mobility assumption could be violated. Substituting the components in (2) for ϵ_{ict} in (4), we observe that a systematic relationship between a teacher's transfer decision and any of these components would violate Assumption 2. In Web Appendix Section A5, I discuss an array of such mechanisms, and I devise and perform tests for the two that are most plausible.

The first mechanism, related to ϕ_{st}, is that teachers systematically transfer toward or away from schools that are about to get better or worse, relative to the school's average quality over the sample period. This might occur, for example, if teachers follow a particularly effective principal when he or she moves from school to school. I test for this mechanism by estimating a model with school-year fixed effects and observing whether schools are disproportionately likely to have

their relatively bad years after transferrers leave or their relatively good years after new transferrers arrive. I do not find evidence that teachers are fleeing declining schools or are flocking to improving schools (see Web Appendix Section A5.1).

The second mechanism, related to the findings of Jackson (2013), is that teachers may systematically transfer toward schools or courses where they are idiosyncratically well-matched. The model presented in Section 2 does not allow for school-teacher or subject-teacher match effects. However, a likelihood ratio test rejects the hypothesis that variation in teacher quality can fully account for the variation in teacher-school match effects, suggesting this scenario could be a concern. For example, a teacher who is adept at using classroom technology may move to a school that provides high-tech classrooms. However, I test for movement toward better school-teacher matches using a subset of schools exhibiting balanced mobility (offsetting flows of teacher transfers in and out), where movement toward better matches is unlikely to bias parameter estimates even if it occurs (see Web Appendix Section A5.2 for further intuition). I find no evidence of such targeted mobility.[18]

Finally, Kramarz, Machin, and Ouazad (2008) show that in a model with both student and school fixed effects, identification also requires that schools and student transfers form a connected graph (with schools as vertices and transferring students as edges). My specification is isomorphic to that of Kramarz, Machin and Ouazad (2008), except that teachers take the place of students, and school-course combinations take the place of schools. Hence, in addition to Assumption 1 (or Assumption 2), identification requires that school-course combinations and teacher transfers form a connected graph, meaning that any two school-course combinations in the network can be linked by some chain of transferring teachers. The existence of a large number of teacher transfers across schools, combined with a substantial fraction of teachers who teach multiple courses or switch courses during the sample, ensures that this rank condition is easily satisfied for a large network of teachers and school-course combinations in the data. Discussion of the connectedness of the

18. In theory, similar concerns could be raised about violations of Assumption 2 among teachers switching subjects within schools. For example, if the only U.S. History teachers observed teaching English 1 at a given school are those pressed into service after a rash of quits among English teachers, these teachers might perform badly relative to the remaining English teachers (compared to their performance among other U.S. History teachers). In this case, we might overestimate the average quality of English teachers at the school and underestimate the average quality of U.S. History teachers (the school-course quality estimates ($\hat{\delta}_{sc}$) for English and U.S. History would be underestimated and overestimated, respectively, so that average test scores in each subject would be fit by the model). In practice, such violations, even if they were to exist, are likely to have minimal impact on estimates of the importance of the allocation of teacher quality for explaining student achievement disparities. First, many such episodes will be offset by teachers switching subjects in the opposite direction (i.e. English teachers being pressed into service in U.S. History in other years). Second, and more importantly, since students are required to take nearly all of these subjects, any biases in the relative quality of teachers teaching different subjects within the school will sum to zero for nearly every student.

network of school-course combinations takes place in Section 4.2. As I will discuss presently, the extent of teacher mobility is important for the precision of parameter estimates.

4 DATA

4.1. Overview

The data, provided by the North Carolina Education Research Data Center, consist of the standardized test scores of all public high school students in North Carolina from 1997-2007 in up to ten subjects, along with a host of student, teacher, and school characteristics.[19] During the sample period, North Carolina offered a standard curriculum with mandatory end-of-course tests for the following subjects: English 1, Econ/Law/Politics, U.S. History, Algebra 1, Algebra 2, Geometry, Biology, Chemistry, Physics, and Physical Science.[20]

The data contain a large number of observable current student[21] and peer[22] inputs that together comprise $\mathbf{X_{ict}}$. Observed prior inputs, $\tilde{\mathbf{Y}}_{\mathbf{i}}^{\mathbf{t-1}}$, include the student's test scores in 7th and 8th grade math and English (standardized by subject-year), along with squares of these test scores, and indicators for missing test scores. Observations were dropped from the sample if fewer than two prior test scores existed. Recall from above that the coefficient associated with each characteristic is allowed to vary with the subject being tested, so that, for example, the impact of a student's 8th grade math test score is allowed to depend on whether the subject currently tested is Algebra 1 or English 1.

To allow a flexible experience profile, teacher experience indicators are created for 6 cells: 0 years of experience, 1 year, 2 years, 3-5 years, 6-11 years, and 12 or more years of experience. $d(x)$ is assumed to equal $d(x')$ for x, x' in the same experience cell.

19. I also use data from 2008-2009 for out-of-sample tests in Section 7.

20. Tests in Physics, Geometry, Chemistry, Physical Science, and Algebra 2 were not introduced until 1999. Also, Econ/Law/Politics was discontinued in 2004, and replaced by Civics and Economics in 2006. U.S. History was not tested in 2004 or 2005.

21. Observable student inputs include the student's race and gender; indicators for parents' education categories, limited English proficiency status, learning disabilities in writing, math, and reading, whether the student is gifted in math or English, grade level (9-12), whether the student is old for his grade, and whether the student is taking the course a year later than his peers at the school. They also include indicators for whether the student intends to attend community college, attend four-year college, or work after high school, as well as indicators for participation in a sport, vocational club, academic club; service club, or arts club.

22. Observable peer/classroom inputs consist of class size, track (honors vs. non-honors), the fraction of the class in each race-gender cell, the fraction of the class in each grade level, the number of gifted students in the class, and class averages of 7th and 8th grade math and reading test scores and their squares.

My empirical strategy requires that student test scores be matched to the teacher who taught the class. Unfortunately, the raw data do not provide an exact match between a test score and the identity of the teacher that taught the class. However, unique classrooms of test scores can be constructed in the test-score-level data, and a list of the classes taught by each teacher in each semester is available in the teacher-level data. Thus, I use a fuzzy matching algorithm to match each teacher-class to a student-class. Since the grade level, race, and gender of each student in the student-class is observed, grade totals and race-gender cell totals can be constructed for the classes in the student-level data and compared to the corresponding grade totals and race-gender cell totals of the classes in the teacher-level data.[23] Test scores from student-level classes whose race, gender, and grade distributions do not closely approximate any teacher-level class in that course in that school are excluded from the analysis that follows. Web Appendix Section A1 describes the implementation of the fuzzy matching algorithm in detail, and provides summary statistics regarding its efficacy. The dataset contains 4.9 million test scores associated with 23,000 teachers in 1,000 public high schools (with 10,000 potential school-subject combinations).

4.2. Connectedness of Schools and Subjects

Recall that identification requires a connected graph of school-course combinations and transferring teachers. Furthermore, estimates are more precise if there are many transferring teachers and if the number of students per teacher is large. Fortunately, the long panel features nearly 3,500 teachers who transfer across schools, and the observed teachers have often taught hundreds of students. I limited the sample of teachers to those who taught at least 20 students. To restrict attention to schools where there is sufficient mobility to plausibly distinguish school and teacher contributions, I impose the following restrictions: (1) each school feature at least 5 teacher transfers, and (2) the final network of school-courses must be two-edge connected, so that any two school-course combinations can be connected using two distinct chains of transfers that do not share any links.[24] These restrictions leave 386 remaining schools and 3,357 remaining school-subject combinations. In fact, the majority of the 386 schools are quite well-connected: 266 of them are connected to the rest of the network by at least 10 transfers. Figure 1a shows the distribution across schools of the number of connecting transfers. Figure 1b shows the distribution of exams administered

23. Students seem to skip ahead or fall behind their grade in one subject fairly often, so that students representing different grades are often observed in the same class.

24. A transferring teacher is defined for these purposes as one who has taught at least 15 students at two different schools.

across teachers in the sample. Figure 1c shows the number of students taught by each transferring teacher at the school at which he/she taught fewer students. The latter figure illustrates that while some teachers have only taught one class at a second school, many others have taught at least 100 students at multiple schools. Table 1 presents a transition matrix describing the patterns of mobility across schools serving different parts of the student background distribution. The rows and columns represent quintiles of the school-level distribution of average student background (as measured by the regression index $X_i\beta + \tilde{Y}_i\alpha$). The top entry in the (i, j)-th cell displays the number of teachers observed teaching in a school in the i-th quintile who moved to a school in the j-th quintile. The bottom entry displays the fraction of teachers ever observed teaching at a school in the i-th quintile who are also observed moving to a school in the j-th quintile. The table shows that while teachers disproportionately move among schools serving similar student populations, there is nonetheless a considerable flow of teachers between schools serving the best- and least-prepared students.[25] Web Appendix Table A3 presents an analogous transition matrix describing the considerable mobility of teachers across subjects. The table reveals that a large fraction of teachers teach multiple courses within their license (i.e. 2-3 distinct science courses, or 2-3 distinct math courses), and that a substantial number even teach courses in multiple fields (e.g. math and history). The fact that switching subjects is so widespread implies that the precision of my teacher quality estimates at a particular school is limited primarily by how well-connected different schools are rather than by the connectedness of subjects within schools.

After dropping students with missing test scores, teachers who only taught at unconnected school-subject combinations, and test scores from classes that were unmatchable, the data still contain 4,016,964 test scores from 822,830 students and 19,826 teachers. Web Appendix Section A2 discusses issues surrounding the choice of test score scale. The raw test scores display no evidence of floor or ceiling effects, and results are robust to applying moderate convex and concave monotonic transformations to the raw scores before standardizing.

25. 37% transfers in the dataset are among schools within the same district. Most districts have a formal procedure by which teachers can request a transfer to a vacancy at another school within the district, so that some of the within-district moves are voluntary. However, superintendents generally reserve the right to trigger involuntary transfers in order to maintain a balance of experience across schools. The data do not indicate whether a transfer was voluntary or not. Because my tests of the exogenous mobility assumption (such as movement away from declining schools or movement toward better school matches) will detect violations regardless of whether the move is triggered by the teacher or the district, I do not further investigate who makes the decision to move. Transfers across districts, like most employer changes, require the mutual agreement of the teacher and the new district.

5 ESTIMATION

From Section 2, the model to be estimated is:

$$Y_{ict} = \tilde{\mathbf{Y}}_{\mathbf{i}}^{\mathbf{t-1}} \alpha_{\mathbf{c}} + \mathbf{X_{ict}} \beta_{\mathbf{c}} + \delta_{sc} + d(ex_{rt}) + \mu_r + \epsilon_{ict} \tag{5}$$

One common approach to estimation taken in the literature is the two-step Empirical Bayes method exemplified by Kane and Staiger (2008). Test scores are regressed on a vector of student observable characteristics in the first stage, then average residuals are formed for each teacher. In the second stage, the year-to-year within-teacher covariance in average student residuals is used as an estimate of the true variance in teacher quality, and each teacher's average residuals are shrunk toward zero using the reliability ratio (where the estimated true variance is used as the signal). However, a key disadvantage of the two-step residual method is that any covariance between student observable characteristics and true teacher quality will necessary load onto the student-level coefficients estimated in the first stage. To the extent that observably superior students disproportionately select superior teachers or good teachers are rewarded with observably superior students, the true variance in teacher quality will be understated.

Because the extent to which already advantaged students receive higher quality teachers is a key focus of the paper, I instead estimate equation (5) in one step via ordinary least squares. While there are over four million test score observations and over twenty thousand parameters, the extreme sparsity of the design matrices makes computation feasible. Between-teacher differences in mean test scores can always be fit perfectly with a full set of teacher effects for any choice of student-level characteristics, school-course effects, and experience effects. Consequently, least squares residuals are minimized by choosing the coefficients on student-level and classroom-level observable characteristics ($\hat{\beta}$ and $\hat{\alpha}$) to best fit the *within-teacher* variation in performance. Similarly, experience effects will adjust to fit the within-teacher growth in performance over time, and school-course effects will adjust to fit the variation in performance across school-subject combinations of teachers who teach in multiple schools or subjects. The teacher effects ($\hat{\mu}$) then adjust to fit the between-teacher variation that remains after removing the predicted regression index of student performance, $X_i\hat{\beta} + \tilde{Y}_i\hat{\alpha} + \hat{\delta}_{sc} + \hat{d}(exp_{rt})$. Because the student-level coefficients, the experience effects, and the school-course effects are only identified using within-teacher variation, they will not reflect any covariance between teacher quality and average student quality, average experience,

or average school-course quality.[26] Thus, the estimated teacher effects will capture the full impact of the teacher's quality, rather than merely the component of teacher quality that could not have been predicted on the basis of student composition, teacher experience, or the school-course in which a teacher teaches.

Analytical asymptotic standard errors are calculated for all parameters. In order to make estimation of the variance-covariance matrix computationally feasible, the calculation is broken down into several pieces and a parametric error components form is imposed in which there are idiosyncratic error components at the test, teacher-year and school-year levels (as specified in (2)). Web Appendix Section A3 describes the details of standard error computation.

Given a limited number of teachers and a limited number of students per teacher, the variance in the estimated distribution of persistent teacher quality, $Var(\hat{\mu})$, will reflect both true variation in μ and variation due to test score measurement error and the other unobserved error components that make up ϵ_{ict}. To distill the true variance in teacher quality, I follow the approach of Aaronson, Barrow, and Sander (2007). I first use the distribution of fixed effect standard errors to estimate the error variance. Then, I subtract the error variance from the estimated fixed effect variance to obtain an estimate of the true variance in μ_r, under the assumption that the true teacher quality and the sampling error are independent. I use the same technique to estimate the true variance in school average teacher quality $\overline{\mu}_s$. Web Appendix Section A4 describes the procedure in detail.

6 RESULTS

6.1. Variance Decomposition

Table 2 contains a decomposition of the variance in student test scores into within-school and between-school components. While only 8.8% of the test score variance is between schools (Row 6, Column 3), the difference in average test scores between a school at the 5th percentile and one at the 95th percentile is nearly a full student-level standard deviation (32nd percentile vs. 68th percentile of the test score distribution). Thus, substantial performance disparities across schools do exist that require explanation.

A closer look at Table 2, however, reveals that average teacher quality and even school quality

26. A symmetric argument made from the perspective of the school-subject effects implies that the teacher effects will be identified using only comparisons within school-subject combinations, with multi-subject teachers acting as bridges to compare teachers from different school-subject combinations, as suggested in Section 3.

in fact have very limited scope to explain average test score differences across schools. Rows 2 and 5 show that the lion's share (90.7%) of the total variance in test scores is due to differences in observable and unobservable student and classroom characteristics and test score measurement error, while Row 7 shows that observable differences in student and classroom characteristics explain three-quarters of the between-school variance. Row 3 indicates that unexplained variation between school-teacher-course-experience category cells accounts for 5 percent of the total variance, suggesting that there is still scope for teacher quality to matter. However, Row 8, labeled "Total School Quality," shows that only 1 percent of the total variance in student test scores is potentially explainable by differences across schools in school quality, average teacher quality, and average teacher experience.

While this may seem surprisingly small, two points are worth noting. First, considerable differences may exist in the quality of the elementary and middle schools attended by students, but these differences will be reflected in differences in average prior test scores $\hat{\alpha}$. High school may be too late to close test score gaps built up through years of unequal family, school, and teacher inputs. Second, comparisons of variances exacerbate differences in the relative importance of various inputs, since variances are measured in units comparable in magnitude to squares of student test scores. Although differences in $\overline{\delta}_s + \overline{\mu}_s + \overline{d(ex)}_s$ across schools explain only 1 percent of the variance, moving from the 5th percentile school to the 95th percentile school in this distribution increases test scores by .33 student-level standard deviations, enough to shift an average student from the 43th percentile to the 57th test score percentile.

6.2. Teacher Experience

The estimated values of the teacher experience profile, $\hat{d}(ex)$, are presented in Table 3. First-year teachers are on average .060 student-level standard deviations worse than second-year teachers, .088 worse than third year teachers, .105 worse than teachers in their fourth through sixth years, .108 worse than teachers in their seventh through twelfth years, and .104 worse than teachers with more than twelve years of experience. This experience profile matches up fairly well with existing estimates from the literature (see Rivkin et al. [2005] or Clotfelter, Ladd, and Vigdor [2007]).

These numbers, combined with the large differences in average teacher experience across schools,[27] may give the false impression that variation in average teacher experience across schools

27. Teachers at a school at the 5th percentile of the average teacher experience distribution have on average eight fewer years of experience than do teachers at a school at the 95th percentile.

has the potential to explain the remaining between-school variation in student test scores. However, the teacher experience differentials across schools are driven in part by differences in the fraction of extremely experienced teachers, for whom the extra few years of experience have little marginal effect on their performance. To account for this, I calculate the average value of effective experience for each school, $\overline{d(ex)}_s$, weighting each teacher-year within the school by the number of students the teacher taught at that school in that year. Panel B of Table 3 displays various quantiles of the distribution of $\overline{d(ex)}_s$. The standard deviation is just .008, and even at the school whose value of $\overline{d(ex)}_s$ puts it at the 1st quantile among schools, the average effective experience of teachers only decreases the average student test score by .020 student level standard deviations, relative to the mean school. This corresponds to a move from the 50th to the 49th percentile for an average student. Simply put, while the first few years of experience do have a significant impact on teacher effectiveness, differences in average teacher experience do not explain the test score gaps we observe across schools in North Carolina. These results may not be surprising when one considers that many school districts have explicit policies allowing involuntary transfers of teachers when one school has too few experienced teachers.

6.3. The True Variance of Teacher Quality

The first row of Table 4 display the raw variances, true variances, and true standard deviations of teacher quality (μ_r) and school average teacher quality ($\overline{\mu}_s$) obtained by pooling across all subjects. These estimates weigh teachers by the number of students taught in the sample, so the standard deviation reported is the standard deviation in teacher quality experienced by randomly selected student-course combinations. While the standard deviation of $\hat{\mu}_r$ is .212, correcting for sampling error leaves an estimate of the true standard deviation in teacher quality of .178 student-level standard deviations. An average student who is assigned a teacher at the 75th percentile of teacher quality can expect to move from the 50th percentile to the 55th percentile, while one who is assigned a teacher at the 95th percentile can expect to move up to the 62th percentile, assuming test scores are distributed normally.[28] This is substantial, and generally in line with most estimates found in the literature at the elementary and middle school level.[29]

I calculate average teacher quality at each school s, denoted $\hat{\overline{\mu}}_s$, by weighting each teacher by

28. This assumption is supported by plots of the data. See Supplementary Figures 1-3

29. Hanushek et al. (2005), for example, find a within-school standard deviation of .22 test score standard deviations. For Kane and Staiger (2008), the standard deviation for English teachers is .17, while the standard deviation for math teachers is .22.

the number of students he/she taught at the school. Applying an analogous measurement error correction, I obtain an estimate of the true between-school standard deviation of teacher quality of .059.[30] The estimate indicates that attending a school whose average teacher quality is in the 75th (95th) percentile moves an average student from the 50th percentile to the 52nd (54th) percentile of the state test score distribution. So while average teacher quality does not vary dramatically across schools, attending a school whose mean teacher quality is well above average can still put a student at a meaningful advantage. Clearly, though, eliminating differences in average teacher quality across high schools would not come close to eliminating test score gaps across schools.

Rows 2-11 of Table 4 present estimates of the variance in teacher quality by subject. By and large, the estimated true standard deviation of teacher quality is fairly consistent across subjects. Specifically, the estimated true standard deviation of teacher quality is .18 in Algebra 1, .18 in Algebra 2, .17 in Geometry, .17 in Biology, .22 in Chemistry, .22 in Physical Science, .20 in Physics, .17 in Econ/Law/Politics, .20 in U.S. History, and only .09 in English. Note, however, that I cannot distinguish the true variation in pedogogical skill from the ability of the test to detect such skill.

Note that much more of the variation in subject-specific teacher quality is between schools than in the estimates that pool across subjects. In other words, much of the between-school variation in teacher quality within each subject seems to be idiosyncratic rather than common to all subjects at the school.

7 Testing for the Existence of Dynamic Tracking

Dynamic tracking of students to teachers represents a key threat to the validity of teacher quality estimates. While my estimates of the variation in average teacher quality across schools should not be sensitive to violations of Assumption 1,[31] estimates of the within-school student-level variation in teacher quality over the span of a high school career could be threatened. Thus, I adapt the

30. I use the delta method to account for correlation in sampling error in $\hat{\mu}_r$ across teachers in the same school when calculating $sd(\overline{\mu_s})$.

31. As discussed in Section 3, the average bias from non-random sorting to teachers within schools is zero by construction at the school level, and thus should approach zero among transferring teachers as the number of transferrers gets large, unless teachers transfer because they are assigned students who are unobservably likely to experience a decline in performance. Furthermore, even if they did so, selection on (negative) value added bias would mean that teachers should have higher average student test-score residuals at the school they move to than the one they left. This is indistinguishable from movement toward higher match quality, which I test and reject in Web Appendix Section A5.2.

test for dynamic tracking within schools introduced by Chetty et al. (2011) to the high school context. The essential insight underlying the test is that every student taking a particular subject in a given year must be taught by some teacher, so non-random sorting among teachers within a subject produces biases that sum to zero within a school-subject combination. Consequently, bias from non-random student sorting can be eliminated by aggregating to the school-subject-year level. As a result, if the estimated variation in teacher effects primarily reflects student sorting on unobservable potential gains rather than true ability to improve student performance, average estimated teacher quality at the subject level should have very little predictive power out of sample. Since the test scores from the years 1997-2007 form the original estimation sample, I use data from the years 2008-2009 to construct estimates of student-weighted average teacher quality for each school-course-year combination, $\overline{\hat{\mu}}_{sct} = \frac{1}{N_{sct}} \sum_{i \in \mathcal{SCT}} \hat{\mu}_{r(i,c,t)}$, as well as estimates of average student test score residuals for each school-course-year combination: $\overline{Z}_{sct} = \frac{1}{N_{sct}} \sum_{i \in \mathcal{SCT}} Y_i - X_{ict}\hat{\beta}_c - \tilde{Y}_i\hat{\alpha}_c - \hat{\delta}_{s(i,t)c} - \hat{d}(exp(r(i,c,t)))$. As Chetty et al. (2011) point out, a univariate regression of the form

$$\overline{Z}_{sct} = \psi_0 + \psi_1 \overline{\hat{\mu}}_{sct} + \xi_{sct} \tag{6}$$

yields an OLS estimate of $\hat{\psi}_1$ that converges in probability to $\frac{Cov(\overline{\hat{\mu}}_{sct}, \overline{Z}_{sct})}{Var(\overline{\hat{\mu}}_{sct})} = \frac{Var(\overline{\mu}_{sct})}{Var(\overline{\hat{\mu}}_{sct})} = \frac{Var(\overline{\mu}_{sct})}{Var(\overline{\mu}_{sct}) + Var(\nu^{\mu}_{sct})}$ under the assumptions of the model (where ν^{μ}_{sct} is sampling error in the estimated average teacher quality for teachers teaching course c in school s at time t). To allow direct comparison of my estimates to the estimates of Chetty et al. (2011) from the middle school level, before running the regression in (6) I shrink the $\overline{\hat{\mu}}_{sct}$ estimates toward zero to form empirical Bayes estimates of $\overline{\hat{\mu}}_{sct}$ via $\mu^{EB}_{sct} = \frac{\hat{Var}(\overline{\mu}_{sct})}{\hat{Var}(\overline{\hat{\mu}}_{sct})} \overline{\hat{\mu}}_{sct}$[32]. Shrinking the estimated school-course-year averages leads to biased but more precise estimates of $\overline{\hat{\mu}}_{sct}$. More importantly, the OLS coefficient on μ^{EB}_{sct} in a regression of

32. Note that for the OLS estimated coefficient to converge to 1 under the assumptions of the model, the EB shrinkage procedure must be performed after averaging among teachers (rather than averaging EB estimates for individual teachers), and must reflect the reliability of estimates of school-course-year averages of teacher quality. Note further that different school-course-years consist of different numbers of teachers whose quality estimates vary in precision depending on the number of students they taught between 1997 and 2007, so that the appropriate shrinkage factor is specific to the school-course-year. Calculating the signal-to-noise ratio for a given school-course-year requires 1) estimating the variance of the true distribution from which $\overline{\mu}_{sct}$ can be considered a draw (denoted $Var(\overline{\mu}_{sct})$), and 2) estimating the variance in the sampling error associated with school-course-year (student-weighted) averages of teacher quality estimates (the precision of each of which depends on the number of students they taught). Both of these estimates require care, and are discussed in Web Appendix Section A4.2.

the form:

$$\overline{Z}_{sct} = \psi_0 + \psi_1 \mu_{sct}^{EB} + \xi_{sct} \tag{7}$$

should now converge in probability to 1 if my estimates of teacher quality are unbiased.

Note that in order to form μ_{sct}^{EB} for a given school-course-year combination in equation (7), I need existing teacher quality estimates $\hat{\mu}_r$ from the 1997-2007 period for each teacher that taught in the chosen school-course-year. Observations associated with new teachers cannot be omitted, since new teachers may systematically be assigned students likely to gain or decline, so that the aggregate bias need not sum to zero among previously observed teachers. Removing all school-course-year combinations from 2008-2009 that feature previously unobserved teachers leaves 1,154 school-course-year observations.[33]

Column 1 of Table 5 displays the results of estimating equation (7). The point estimate of $\hat{\psi}_1$ is .832, with a standard error of .087. While the aggregated teacher quality estimates clearly have very good predictive power out-of-sample, a t-test rejects equality with 1 at the 10 percent level (though not at the 5 percent level).

To verify that dynamic sorting of students unobservably likely to improve to particular teachers is indeed what is driving $\hat{\psi}_1$ below 1, I follow Jackson (2012) and exploit the fact that out-of-sample differences in student residuals among teachers teaching in the same school-subject-year should reflect both differences in true quality as well as persistent patterns of dynamic sorting within schools and courses. Specifically, let $\tilde{Z}_{rsct}$ represent deviations in teacher r's average student residuals in school s and course c at year t relative to the mean student residuals in that school-course-year combination:

$$\tilde{Z}_{rsct} = \frac{1}{N_{rsct}} \sum_{i \in \mathcal{RSCT}} Y_i - X_{ict}\hat{\beta}_c - \tilde{Y}_i\hat{\alpha}_c - \hat{\delta}_{s(i,t)c} - \hat{d}(exp(r(i,c,t))) - \overline{Z}_{sct} \tag{8}$$

Similarly, let $\tilde{\hat{\mu}}_r^{sct}$ represent deviations in teacher r's estimated quality from the estimated (student-weighted) average among all teachers from the same school-course-year:

$$\tilde{\hat{\mu}}_r^{sct} = \hat{\mu}_r - \frac{1}{N_{sct}} \sum_{i \in \mathcal{SCT}} \hat{\mu}_{r(i,c,t)} \tag{9}$$

33. Note that many of the previously unobserved teachers are not novice teachers, but merely teachers that have not previously taught in a course for which there was a statewide standardized test.

Consider regressing $\tilde{Z}_{rsct}$ on the appropriately shrunken empirical Bayes estimate[34] of $\tilde{\hat{\mu}}_r^{sct}$, denoted $\tilde{\mu}_{rsct}^{EB}$:

$$\tilde{Z}_{rsct} = \zeta_0 + \zeta_1 \tilde{\mu}_{rsct}^{EB} + \tilde{\xi}_{rsct} \tag{10}$$

If the differences in performance among teachers from the same school-course-year are driven by a combination of persistent differences in true teacher quality and persistent differences in ability to attract students whose performance is likely to exceed what could be predicted based on observables, then $\hat{\zeta}_1 \rightarrow_p 1$.

Column 2 of Table 5 displays the results of estimating equation (10). The point estimate of $\hat{\zeta}_1$ is 1.10, with a standard error of .082, which is not statistically significantly different than 1 at either the 5% or 10% level. These results confirm that some degree of dynamic tracking is likely to be occurring. However, as discussed in Section 3, while dynamic tracking may be slightly inflating my estimates of the true within-school variance in teacher quality, it is unlikely to affect my estimates of the true between-school variance in teacher quality. Furthermore, dynamic tracking need not bias the estimates presented below of the extent to which disadvantaged students are receiving the relatively inferior students within their schools. It will only do so if observably disadvantaged students are systematically more or less likely to be assigned to the teachers within their schools who are relatively effective at receiving students who are unobservably more likely to improve. Thus, the basic results regarding the contribution of teacher allocation to student test score inequality are unlikely to be substantially affected by the moderate amount of dynamic tracking suggested by the out-of-sample regression results.

8 Student-Level Variance in Average Teacher Quality

While the results indicate that differences in overall average teacher quality across schools are modest, the sizable within-school variance in teacher quality may still contribute substantially to inequality if some students are consistently assigned to poor teachers in course after course, relative to their school's average. This could be the result of pure bad luck, but could also occur

34. See Web Appendix Section A4.2 for a description of the shrinkage procedure for this specification.

systematically if students are choosing course tracks and the best teachers within a school tend to be assigned to the honors track.[35] On the other hand, if each student is taught by offsetting combinations of good and bad teachers, even a substantial amount of variation in teacher quality at a school need not lead to sizable differences across students in the quality of teaching they receive. To examine the variation in student-level teaching quality within schools, I first calculate the average estimated teacher quality across courses for each student who took tests in at least five different courses, relative to the overall average teacher quality at the student's school. I denote this measure by $\hat{\overline{\mu}}_i - \hat{\overline{\mu}}_s$.[36] Using the delta method to calculate standard errors for each student's average teacher quality, $\sigma_i^{\overline{\mu}}$, I can estimate the variance in student-level teacher quality within schools as:

$$Var(\tilde{\overline{\mu}}_i) = Var(\overline{\mu}_i - \overline{\mu}_s) = Var(\hat{\overline{\mu}}_i - \hat{\overline{\mu}}_s) - (1/I)\sum_i (\sigma_i^{\overline{\mu}})^2. \tag{11}$$

Among students who took at least five tests, a one standard deviation increase in average teacher quality (relative to the overall school average) corresponds to an increase in average teacher quality of .048 test score standard deviations. In other words, a student whose average teacher is at the 10th (90th) percentile of the student-level average teacher quality distribution will have his expected test scores in each course reduced (increased) by an average of .062 standard deviations, solely by virtue of the teachers he was assigned at his school. This is enough to shift an average student from the 50th to the 48th (52rd) test score percentile. Thus, assignment of teachers to students within schools contributes nearly as much to the test-score variation across students as does variation in average teacher quality across schools.

However, to the extent that this variation in student-level teacher quality is attributable to simple good and bad luck, it seems difficult to remedy. Thus, I also estimate what the student-level variance in teacher quality would be if students were randomly assigned to their teachers, subject to the important constraint that all students have to take each subject. After all, in the extreme case of a small school with only one biology teacher, one chemistry teacher, and one physics teacher, there may be considerable variation in the quality of science instruction across these teachers, but each student at this school will have the same three science teachers.[37]

35. Note that such non-random assignment of students to teachers need not bias the estimates of teacher quality if the students assigned to the best teachers are *predictably* superior based on prior test scores and the average prior test scores of those in their classes.

36. The results are similar if I condition on six or seven tests.

37. Note that "across-subject variation" in this context refers to schools who have, say, relatively good algebra teachers compared to the state's average algebra teacher, but relatively poor biology teachers compared to the state's

For each student I construct a set of feasible paths of teachers that the student could have experienced, given the sets of teachers that were teaching the subjects the student took when he took them at his school. Then, I randomly select a path of teachers for each student from these student-specific sets of feasible paths, and calculate the variance in average simulated within-school teacher quality across students. After repeating this simulation 100 times and averaging across simulated samples, I find that the across-student standard deviation in teacher quality under random assignment is .046 test score standard deviations.[38] Thus, within-school variation in the average teacher quality experienced across students does contribute to performance heterogeneity, but the variance in average teaching quality we observe is essentially equal to what we would expect under random assignment.

9 The Impact of Teacher Inputs on Achievement Inequality

Given knowledge of the underlying distribution of teacher quality, in this section I examine the extent to which the existing allocation of students to teachers is harming disadvantaged students. First, I look at whether disadvantaged students are more likely to receive their schools' relatively ineffective teachers. Second, I consider the extent to which the average teaching quality is lower at schools that disproportionately serve underprivileged students. Finally, I provide an aggregate measure of the contribution of teacher inputs to achievement gaps between advantaged and disadvantaged students.

average biology teacher.

38. I actually use two different methods for constructing feasible paths of teachers for each student. The first method includes any permutation of teachers that taught the appropriate subjects at the appropriate times at the appropriate high school. However, this may overestimate the range of teaching possibilities available to the student if there are scheduling conflicts (i.e. the student took English and chemistry in the same year, and one of the English teachers taught at the same time as one of the chemistry teachers, making this combination of teachers infeasible). Thus, to get a lower bound on the variance in average teacher quality across students under random assignment, I also performed the analysis using only paths of teachers that were actually experienced by some student who took the same sequence of courses in the same years as the student in question. This clearly understates the variance under random assignment because some feasible combinations of teachers may not have been actually chosen by any one student. The results were not sensitive to the method chosen, suggesting that either scheduling conflicts were rare, or most feasible paths were taken by some student.

9.1. Are Underprivileged Students Systematically Assigned to Classes with Inferior Teachers?

One way to interpret the results from Section 8 is that the existing mechanisms for allocating teachers to classes are not contributing to inequality in test score performance. However, it is still possible that students in lower tracks are systematically receiving lower quality teachers, and that the impact of such an imbalance on the student-level variance that I estimated is being masked by some other feature of the teacher allocation mechanism that is reducing variance in average teacher quality among students on the same track.

Consequently, I also examine more directly whether students with observable characteristics that predict lower performance tend to receive their schools' relatively ineffective teachers over the course of their high school careers. More specifically, I first form an index of student background, $(X_{ict}\hat{\beta} + \tilde{Y}_i\hat{\alpha})$, by weighting student characteristics by how well they predict high school test score performance.[39] Then, for each student among the bottom 10% of this index, I compare the average estimated quality of the teachers that actually taught the student with the average estimated quality among the set of teachers who taught the subjects the student took in the years they took them in their schools. I find that such students received teachers that were .025 test score standard deviations less effective than the average teacher across the set of feasible paths of teachers available to them. Students in the top 10% of the index received teachers that were .025 better than the average teacher they could have expected under random assignment, given the teachers teaching in their subjects at the time. I find that black students received teachers who were .014 test score standard deviations less effective then they could have expected under random assignment. If instead I use as the baseline the average teacher along only those feasible paths of teachers that some other student at their school actually took during the years each student was taking his/her respective courses, students in the bottom 10% of the background index received teachers who were .013 test score standard deviations below the average among feasible paths. Students in the top 10% receive teachers who are .013 above the mean among feasible paths satisfying this restriction, while black students receive teachers who are .011 worse than the mean among feasible paths. Thus, I do find modest evidence that disadvantaged students are systematically more likely to be taught by the relatively ineffective teachers at their schools.

39. Note that X_{ict} also includes the average prior test scores of a student's classmates, so that this index also partially reflects the kinds of peers they interact with.

9.2. What Kinds of Teachers Teach at Schools that Disproportionately Serve Underprivileged Students?

The estimated parameter distributions presented in Section 6 allow us to examine whether the schools disproportionately serving underprivileged students generally offer inferior teaching. To this end, Table 6 provides the average values of $\overline{\hat{\mu}}_s$ among schools in the top quartile and bottom quartile of a set of salient measures of average student background. The signs for school average teacher quality (Columns 3 and 4) generally conform to expectations: schools whose students have low prior test scores have below average teacher quality, as do schools with a high percentage of students who are eligible for free lunch, and schools with a high fraction of black students. However, the magnitudes are small, in keeping with the general finding that very little of the variance in teacher quality is between schools. These small differences also mirror the findings of Sass et al. (2010) at the elementary school level.

The last row presents results for my most comprehensive measure of average student background, the average value of the index $X_{ict}\beta + \tilde{Y}_i\alpha$. I find that high schools whose average indices across students place them in the bottom quartile of schools have teachers who are .049 student level standard deviations below average, while those in the top quartile have teachers who are .027 standard deviations above average. Thus, much of the between-school variation in teacher quality can be predicted based on the composition of schools' student bodies: -.049 is at the 20th percentile of the school average teacher quality distribution, while .027 is at the 67th percentile of the school average teacher quality distribution (assuming normality).

9.3. The Aggregate Contribution of Teacher Inputs to Inequality of Opportunity

While Table 6 provides a good sense of the inputs provided by the schools most heavily populated by disadvantaged students, even these schools serve a mix of under-supported and well-supported students. Thus, to address more directly the contribution of unequal teacher inputs to disparities in achievement, I examine the typical allotments of these inputs received by particular subpopulations of students. The subpopulations I consider are black students, white students, and students in the bottom and top deciles and quartiles of the student background index, $X_{ict}\beta + \tilde{Y}_i\alpha$. Table 7 displays for each subpopulation the average values of school-average teacher quality and

effective teacher experience along with within-school deviations in teacher quality and effective experience from the school mean.

I find that students in the bottom decile of the student background index attend schools where teachers are on average .011 test score standard deviations below the state mean. They are also assigned to classrooms with teachers that are on balance .029 test score standard deviations below their schools' averages. Below-average teacher experience among their teachers reduces their test scores by an additional .003 test score standard deviations. The combined teaching environment they experience (as measured by $\overline{\mu}_i + \overline{d(ex)}_i$) lowers their test scores by .044 standard deviations, relative to a typical teaching environment in North Carolina. Students in the top decile go to schools where teacher quality is on average .008 above the state mean. They are assigned teachers that are .023 standard deviations above their schools' averages. Once teacher experience has been factored in, the teaching environment they enjoy raises their test scores by .035 standard deviations. However, the cumulative difference in the teacher inputs the two groups receive can only account for .079 of the 2.58 standard deviation difference in their test scores (3.1%). Thus, the way teacher inputs are allocated in North Carolina is only making a marginal contribution to what was already a massive difference in high school achievement between the least supported and best supported students.

The comparison between races reveals a similar pattern. Black students attend schools with slightly below average teachers and receive slightly below average teachers within those schools, so that the overall high school teaching environment they face lowers their test scores by .022 standard deviations relative to the average student. Differences in typical high school teaching environments can account for 4.2% of the .75 standard deviation black-white test score gap. Thus, teacher allocation in North Carolina high schools is making a modest but non-negligible contribution to racial achievement inequality.

10 CONCLUSION

In contrast to the horror stories recounted in the popular media in which the least privileged students attend disorganized schools with ineffective teachers, I find instead that quality teaching is fairly equitably distributed within and across public high schools in North Carolina. Disadvantaged students, whether indicated by race, free lunch eligibility, or low values of the predictive index $X_{it}\beta + \tilde{Y}_i\alpha$, do tend to be exposed to slightly inferior high school teachers across several measures

of student disadvantage. However, the contribution of this source of inequality to overall test score achievement gaps is small, suggesting that high school may be too late to intervene on behalf of underprivileged students.

Why don't we see stronger sorting of teachers to schools? One explanation may be the limited financial incentive a good school can offer, since the bulk of public school teachers' salaries are funded by the state in North Carolina, and all teachers in the same district with the same credentials and experience are paid the same salary. Alternatively, the limited assortative matching of effective teachers to desirable schools may partly reflect inadequate information by such schools at the time of hiring, since previous research suggests that teacher characteristics that are easily observable at the time of hiring are weak indicators of teacher quality (Clotfelter, Ladd, and Vigdor [2007]; Rockoff et al. [2008]). Such information scarcity is exacerbated by the notorious difficulty administrators have in firing underperforming teachers (even in a state without collective bargaining), so that hiring mistakes may be difficult to rectify.

Since transfer patterns provide only faint evidence of an underlying job ladder (see Web Appendix Section A6), a third possibility is that teachers hold weak or horizontal preferences among schools, so that the notion of universally "desirable" schools is inaccurate, even if preferences for particular school characteristics may be vertical (e.g. neighborhood crime rates).[40]

Another explanation may lie in the fact that teachers are hired by districts rather than schools, so that for the 37% of transfers that are occurring within districts, transfer patterns may more closely reflect the preferences of district administrators than those of teachers. In this case, within-district job desirability would not be reflected in teacher transfers. Moveover, if administrators value equality of opportunity and have sufficient knowledge of experienced teachers' relative qualities when transfer opportunities arise, their teacher reallocation decisions may actually be contributing to the relative teaching equality across schools (at least within districts).

While differences in average teacher quality explain a small fraction of performance gaps across schools, I do find that teachers matter, even at the high school level. Within-school variation in teacher quality accounts for a non-trivial fraction of the within-school test score variance. Assignment to a teacher who is one standard deviation above average raises a student's expected test score by .178 student-level standard deviations, enough to shift an average student from the 50th to the

40. Research by Boyd et al. (2005) suggests that distance from home is perhaps the strongest factor in teacher location decisions. If teachers are drawn from all over the state, this finding may partly explain disagreement among teachers in preferences over schools. Evidence that teachers' quit rates and location decisions do respond systematically to some school characteristics can be seen in Goldhaber, Gross and Player (2007) and Jackson (2009).

57th percentile of the state test score distribution.

Given the sizable variation in teacher quality within schools, I explore the contribution of variation in the average teacher quality experienced across courses in high school to performance differences among students attending the same schools. While I find that which teachers a given student happens to receive has a modest but non-negligible impact on his overall performance in high school, the variation in average teaching quality experienced across students only barely exceeds what would be produced by random assignment of students to teachers within a school. I find that disadvantaged students are only slightly less likely to receive the relatively effective teachers at their high schools.

A caveat merits mention. The distribution of teacher quality characterized by this paper reflects the equilibrium that existed in North Carolina between 1997 and 2007. If we change the mechanisms by which teachers are recruited and evaluated, the content of the curricula upon which the subject tests are based, or the manner in which parents and students sort into schools, we should expect to move to a new equilibrium that exhibits a distinct joint distribution of school, student, and teacher quality. For this reason, we must be cautious about generalizing these results to other states, grade levels, or outcomes.

DEPARTMENT OF ECONOMICS AND SCHOOL OF INDUSTRIAL AND LABOR RELATIONS
CORNELL UNIVERSITY

REFERENCES

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

Abowd, John, Francis Kramarz, and David Margolis (1999) 'High Wage Workers and High Wage Firms.' *Econometrica* 67(2), 251–333

Boyd, Donald, Hampton Lankford, Susanna Loeb, and James Wyckoff (2005) 'The Draw of Home: How Teachers Preferences for Proximity Disadvantage Urban Schools.' *Journal of Policy Analysis and Management* 24(1), 113–132

Boyd, Donald, Pam Grossman, Hampton Lankford, Susanna Loeb, and James Wyckoff (2007) 'Who Leaves? Teacher Attrition and Student Achievement.' NBER Working Paper 14022, National Bureau of Economic Research, Inc

Chetty, Raj, John Friedman, and Jonah Rockoff (2011) 'The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood.' NBER Working Paper 17699

Clotfelter, Charles, Helen Ladd, and Jacob Vigdor (2006) 'Teacher-Student Matching and the Assessment of Teacher Effectiveness.' *Journal of Human Resources* 41(4), 778–820

___ (2007) 'Teacher Credentials and Student Achievement: Longitudinal Analysis with Student Fixed Effects.' *Economics of Education Review* 26(6), 673–682

Goldhaber, Daniel, Bethany Gross, and Daniel Player (2007) 'Are Public Schools Really Losing Their Best? Assessing the Career Transitions of Teachers and Their Implications for the Quality of the Teacher Workforce.' Working Paper 12, The Urban Institute. National Center for Analysis of Longitudinal Data in Education Research

Hanushek, Eric, John Kain, Daniel O'Brien, and Steven Rivkin (2005) 'The Market for Teacher Quality.' NBER Working Paper 11154, National Bureau of Economic Research, Inc

Harris, Douglas, and Tim R. Sass (2006) 'Value-Added Models and the Measurement of Teacher Quality.' Unpublished Manuscript

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

___ (2012) 'Do High School Teachers Really Matter?' NBER Working Paper 17722, National Bureau of Economic Research, Inc

___ (2013) 'Match Quality, Worker Productivity, and Worker Mobility: Direct Evidence from Teachers.' *Review of Economics and Statistics*

Kane, Thomas, and Douglas Staiger (2008) 'Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation.' NBER Working Paper 14607, National Bureau of Economic Research, Inc

Kane, Thomas, Jonah Rockoff, and Douglas Staiger (2008) 'What Does Certification Tell Us about Teacher Effectiveness: Evidence from New York City.' *Economics of Education Review* 27(6), 615–631

Koedel, Cory (2010) 'Teacher Quality and Dropout Outcomes in a Large Urban School Distict.' *Journal of Urban Economics* 64(3), 560–572

Koedel, Cory, and Julian Betts (2010) 'Value Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation.' *Education Finance and Policy* 5(1), 54–81

Kramarz, Francis, Stephen Machin, and Amine Ouazad (2008) 'What Makes a Test Score? The Respective Contributions of Pupils, Schools and Peers in Achievement in English Primary Education.' Discussion Paper No. 3866, Institute for the Study of Labor in Bonn

Lankford, Hampton, Susanna Loeb, and James Wyckoff (2002) 'Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis.' *Educational Evaluation and Policy Analysis* 24(1), 37–62

Lise, Jeremy, Costas Meghir, and Jean-Marc Robin (2002) 'Matching, Sorting, and Wages.' Unpublished Manuscript

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

Lopes de Melo, Rafael (2009) 'Sorting in the Labor Market: Theory and Measurement.' Unpublished Manuscript

Meghir, Costas, and Steven Rivkin (2010) 'Econometric Methods for Research in Education.' NBER Working Paper 16003. Prepared for the Handbook of Education., National Bureau of Economic Research, Inc

Pop-Eleches, Christian, and Miguel Uruiola (2011) 'Going to a Better School: Effects and Behavioral Responses.' NBER Working Paper 16886, National Bureau of Economic Research

Rivkin, Steven, Eric Hanushek, and John Kain (2005) 'Teachers, Schools, and Academic Achievement.' *Econometrica* 73(2), 417–458

Rockoff, Jonah (2004) 'The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data.' *American Economic Review: Papers and Proceedings of the One Hundred Sixteenth Annual Meeting of the American Economic Association* 94(2), 247–252

Rockoff, Jonah, Brian Jacob, Thomas Kane, and D. Staiger (2008) 'Can You Recognize An Effective Teacher When You Recruit One.' NBER Working Paper 14485, National Bureau of Economic Research, Inc

Rothstein, Jesse (2010) 'Teacher Quality in Education Production: Tracking, Decay, and Student Achievement.' *Quarterly Journal of Economics* 125(1), 175–214

Sass, Tim, Jane Hannaway, Zeyu Xu, David Figlio, and Li Feng (2010) 'Value Added of Teachers in High-Poverty Schools and Lower Poverty Schools: Implications for Research, Policy and Management.' Working Paper 52, National Center for Analysis of Longitudinal Data in Education Research

Steele, Jennifer, Richard Murnane, and John Willet (2010) 'Do Financial Incentives Help Low-Performing Schools Attract and Keep Academically Talented Teachers? Evidence from California.' *Journal of Policy Analysis and Management* 29(3), 451–478

Todd, Petra, and Kenneth Wolpin (2003) 'On the Specification and Estimation of the Production Function for Cognitive Achievement.' *The Economic Journal* 113, F3–F33

Xu, Zeyu, Jane Hannaway, and Colin Taylor (2011) 'Making A Difference? The Effects of Teach for America in High School.' *Journal of Policy Analysis and Management* 30(3), 447–469

11 TABLES AND FIGURES

TABLE 1: THE PATTERN OF TEACHER MOBILITY BETWEEN SCHOOLS IN DIFFERENT QUINTILES OF THE DISTRIBUTION OF AVERAGE STUDENT BACKGROUND (AS MEASURED BY $X_{it}\beta + \tilde{Y}_i\alpha$)

	School Quintiles				
	1st	2nd	3rd	4th	5th
1st	228 0.052	195 0.044	175 0.040	146 0.033	121 0.027
2nd	121 0.027	146 0.032	196 0.043	214 0.047	159 0.035
3rd	93 0.020	153 0.033	171 0.036	213 0.045	169 0.036
4th	75 0.014	138 0.026	193 0.036	209 0.039	210 0.039
5th	61 0.013	83 0.017	114 0.024	171 0.036	229 0.048

Note: Rows and columns are quintiles of the school-level average of the predictive index of student performance across schools.
The top entry is the number of teachers who moved from schools in the row quintile to schools in the column quintile of school-average student predicted quality.
The bottom entry is the fraction of teachers who ever taught in a school in the row quintile who moved to a school in the column quintile.

TABLE 2: VARIANCE DECOMPOSITION OF STUDENT TEST SCORES

	Variance Component	Variance (1)	Standard Deviation (2)	Fraction of Total Var. (3)
(1)	Total: $Var(Y_{ist})$	.941	.970	–
	Components:			
(2)	Student Background $Var(X_{ict}\beta_c + \tilde{Y}_i^{t-1})$	.560	.748	.596
(3)	Effective School and Teacher Quality $Var(\lambda_{srcj})^*$	.052	.228	.055
(4)	Cov(Stu. Background, Eff. Sch./Tch. Qual.) $2 * Cov(X_{ict}\beta_c + \tilde{Y}_i^{t-1}, \lambda_{srj})$	.034	–	.037
(5)	Idiosyncratic Test Score Error $Var(\epsilon_{ict})$	.291	.540	.310
(6)	Between School Total: $Var(\overline{Y}_s)$	.083	.288	.088
	Components:			
(7)	School Average Student Background $Var(\overline{X_s\beta_c} + \overline{\tilde{Y}}_s^{t-1})$	.062	.249	.066
(8)	Total School Quality $Var(\overline{\lambda}_s)^{**}$	.010	.099	.010
(9)	Cov(Avg. Stu. Background, Total Sch. Qual.) $2 * Cov(\overline{X_s\beta_c} + \overline{\tilde{Y}}_s^{t-1}, \overline{\lambda}_s)$	.012	–	.013

Note: $^*\lambda_{srcj}$ is the mean unpredicted test score of students taught by teacher r in school s in course c while the teacher was in experience cell j. $\lambda_{srcj} = \delta_{sc} + \mu_r + d(ex_{rt})) + \omega_{srcj}$. Thus, $Var(\lambda_{srcj})$ consists of the combined contributions of school-course quality, teacher quality, teacher experience, and the component of the idiosyncratic error that is between school-teacher-course-experience cells. Note that "Student Background" includes the impact of classroom peers, since average observable characteristics of classmates are elements of X_{ict}.

TABLE 3: THE IMPORTANCE OF TEACHER EXPERIENCE FOR IMPROVING TEST SCORES AND FOR EXPLAINING STUDENT PERFORMANCE GAPS ACROSS SCHOOLS

Panel A: The Impact of Teacher Experience on Student Standardized Test Scores

	Years of Experience					
Ex	0	1	2	3-5	6-11	12+
$\hat{d}(Ex)$	0 (0)	.060 (.004)	.088 (.004)	.105 (.004)	.108 (.005)	.104 (.007)

Panel B: The Distribution of Average Teacher Effective Experience ($\overline{\hat{d}(ex)}_s$) Across Schools

	Quantile of $\overline{\hat{d}(ex)}_s$					
	1st %	5th %	25th %	75th %	95th %	99th %
Test Score SDs	-0.020	-0.010	-0.003	0.004	0.009	0.011
Test Score Percentile (Avg. Student)	49.2%	49.6%	49.9%	50.2%	50.4%	50.5%

Note: The distribution of $\overline{\hat{d}(ex)}_s$ has been re-normalized to have a median of zero, so that the table entries in Row 1 of Panel B reflect the impact on expected test scores of attending a school whose average teacher effective experience places it at the k-th quantile, relative to attending a school with median effective experience among its teachers. Row 2 displays the test score percentile that the median student (50th percentile) would obtain if they instead experienced the teacher experience distribution of the k-th quantile school.

TABLE 4: RAW AND ERROR-ADJUSTED VARIANCES IN TEACHER QUALITY AND SCHOOL-AVERAGE TEACHER QUALITY (μ_r, $\overline{\mu}_s$): AGGREGATE AND SUBJECT-SPECIFIC

	Total (μ_r)			Between-School ($\overline{\mu}_s$)		
Parameter	Raw Var.	True Var.	True Std.	Raw Var.	True Var.	True Std.
All Subj. (Stu. Wgt)	.045 (.002)	.031 (.002)	.178 (.006)	.006 (.001)	.003 (.001)	.059 (.004)
Algebra 1	.045 (.001)	.034 (.001)	.184 (.003)	.016 (.001)	.011 (.001)	.102 (.006)
Algebra 2	.043 (.002)	.034 (.002)	.184 (.004)	.017 (.002)	.012 (.002)	.107 (.007)
Geometry	.037 (.001)	.028 (.001)	.166 (.004)	.017 (.002)	.011 (.002)	.104 (.007)
Biology	.042 (.001)	.029 (.001)	.170 (.004)	.021 (.002)	.012 (.002)	.111 (.009)
Chemistry	.063 (.003)	.048 (.003)	.220 (.007)	.033 (.004)	.022 (.004)	.149 (.012)
Physical Science	.060 (.005)	.046 (.005)	.215 (.012)	.039 (.006)	.028 (.005)	.168 (.016)
Physics	.072 (.002)	.039 (.002)	.196 (.005)	.024 (.002)	.016 (.002)	.125 (.008)
Civics/ELP	.045 (.002)	.030 (.001)	.172 (.004)	.024 (.003)	.014 (.002)	.117 (.009)
U.S. History	.057 (.002)	.041 (.002)	.202 (.005)	.030 (.003)	.019 (.003)	.139 (.012)
English 1	.024 (.001)	.008 (.001)	.092 (.005)	.015 (.002)	.005 (.002)	.070 (.013)

Note: "All Subj. (Stu. Wgt)" weights each subject by its fraction of all observed student-subject combinations.

Approximate standard errors are in parentheses. They were obtained using bootstrap samples from the combinations of $\{\hat{\mu}, sd(\hat{\mu})\}$ or $\{\hat{\overline{\mu}}, sd(\hat{\overline{\mu}})\}$ estimates. They are likely to be underestimates, since the individual parameter estimates are held fixed across bootstrap samples. Re-estimating the model (along with calculating analytical standard errors for individual parameters) for each bootstrap sample was computationally infeasible.

TABLE 5: OUT-OF-SAMPLE TESTS FOR PERSISTENCE OF TEACHER QUALITY AND DYNAMIC TRACKING USING DATA FROM 2008-2009

	Specification/Level of Observation	
	School-Course-Year Averages	Teacher-Specific Deviations from School-Course-Year Averages
	(1)	(2)
$\hat{\psi}_1$	.832 (.087)	–
$\hat{\zeta}_1$	–	1.10 (.082)
N	1154	1206

Note: "School-Course-Year Averages" refers to a specification in which school-course-year averages of student residuals from the years 2008-2009 are regressed on school-course-year averages of predicted teacher quality (based on shrunken estimates from the years 1997-2007): $\overline{Z}_{sct} = \psi_0 + \psi_1 \mu^{EB}_{sct} + \xi_{sct}$

"Teacher-Specific Deviations from School-Course-Year Averages" refers to a specification in which teacher-school-course-year specific deviations of average student residuals from school-course-year means from the years 2008-2009 are regressed on teacher-school-course-year specific deviations of predicted teacher quality from school-course-year means (based on shrunken estimates from the years 1997-2007): $\tilde{Z}_{rsct} = \zeta_0 + \zeta_1 \tilde{\mu}^{EB}_{rsct} + \tilde{\xi}_{rsct}$

TABLE 6: AVERAGE TEACHER QUALITY AMONG SCHOOLS IN THE TOP QUARTILE VERSUS BOTTOM QUARTILE OF VARIOUS STUDENT CHARACTERISTICS

	Mean Student Characteristic		Mean Teacher Quality ($\hat{\bar{\mu}}_s$)	
	Bottom	Top	Bottom	Top
Mean 8th Grade Math Score	-.444	.427	-.027 (.007)	.031 (.007)
Fraction Black	.051	.600	.020 (.008)	-.042 (.008)
Fraction Eligible for Free Lunch	.139	.542	.020 (.005)	-.044 (.010)
Stu. Backgr. Index ($X_i\hat{\beta} + \tilde{Y}_i^{t-1}\hat{\alpha}$)	-.393	.368	-.049 (.009)	.027 (.006)

Note: Mean Student Characteristic is the average value of the student characteristic associated with a given row among the schools in either the bottom or top quartile of schools sorted by their values of that characteristic.

Mean Teacher Quality is the average value of estimated average teacher quality ($\hat{\bar{\mu}}_s$) among schools in either the top or bottom quartile of schools sorted by their values of the student background measure associated with a given row.

Stu. Backgr. Index is an index of student background composed of the predicted test score based solely on the student's current observable characteristics and test scores collected prior to high school.

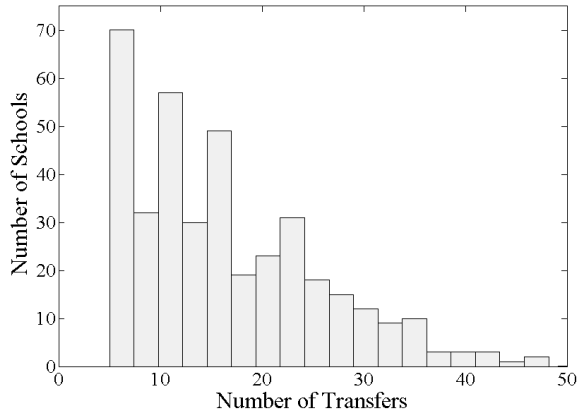
TABLE 7: AVERAGE SCHOOLING INPUTS AND OUTCOMES AMONG SELECTED SUBPOPULATION OF STUDENTS, IN TEST SCORE STANDARD DEVIATIONS (BASELINE SPECIFICATION)

	Average Value of Input or Outcome Among Subpopulation					
Student Subpopulation	Test Score $(\overline{Y}_i)$	Within Sch. Tch. Qual. $(\overline{\hat{\mu}}_i - \overline{\hat{\mu}}_s)$	Between Sch. Tch. Qual. $(\overline{\hat{\mu}}_s)$	Within Sch. Tch. Exper. $(\overline{\hat{d}(ex)}_i - \overline{\hat{d}(ex)}_s)$	Between Sch. Tch. Exper. $(\overline{\hat{d}(ex)}_s)$	Total Teacher Contribution $(\overline{\hat{\mu}}_i + \overline{\hat{d}(ex)}_i))$
Student Background Index ($X_{it}B + \tilde{Y}_i^{t-1}\alpha$)						
Bottom 10%	-1.25	-.029	-.011	-.002	-.001	-.044
Bottom 25%	-.935	-.020	-.008	-.002	-.001	-.032
Top 25%	.975	.019	.007	.003	.000	.029
Top 10%	1.33	.023	.008	.004	.000	.035
Race						
White	.224	.004	.005	.000	.001	.010
Black	-.522	-.009	-.010	-.001	-.002	-.022

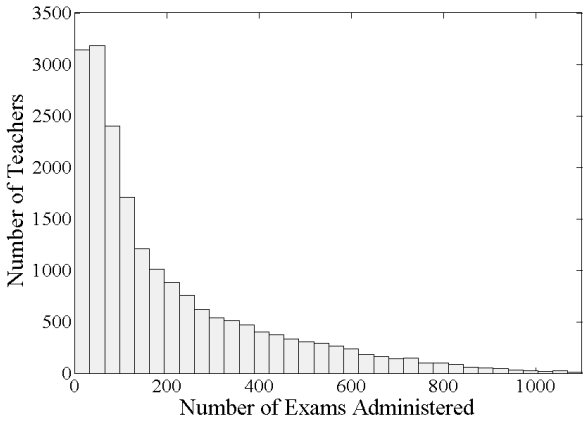
Note: "Bottom 10%" refers to the set of students whose value of the background index $X_{it}B + \tilde{Y}_i^{t-1}\alpha$ places them below the 10th quantile of the distribution of the index among all students.
"Total Teacher Contribution" consists of the combined contributions of assignments to higher/lower quality teachers within schools, higher/lower average teacher quality at the schools attended by the chosen subpopulation, assignments to more/less experienced teachers within schools, and higher/lower average teacher experience at the schools attended by the chosen subpopulation: $(\overline{\hat{\mu}}_i + \overline{\hat{d}(ex)}_i)$.

FIGURE 1: A GRAPHICAL DEPICTION OF THE NETWORK OF TEACHER TRANSFERS

(a) DISTRIBUTION OF THE NUMBER OF TRANSFERRERS ACROSS SCHOOLS



(b) DISTRIBUTION OF THE NUMBER OF EXAMS ADMINISTERED ACROSS TEACHERS



(c) DISTRIBUTION OF MIN(TOTAL STUDENTS$_1$, TOTAL STUDENTS$_2$) FOR TRANSFERRING TEACHERS