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 examines the extent to which the allocation of teachers within and across public high schools is contributing to inequality in student test score performance. Using ten years of administrative data from North Carolina public high schools, I estimate a flexible education production function in which student achievement reflects student inputs, teacher quality, school quality, and a school-specific scaling factor that allows the impact of teaching quality to vary across schools. The existence of nearly 3,000 teacher transfers, combined with a testable exogenous mobility assumption, allows separate identification of each teacher's quality from both school quality and school sensitivity to teacher quality. I find that teaching quality is surprisingly equitably distributed both within and across high schools. Schools predominantly serving underprivileged students employ teachers who are only slightly below average, and most students receive a mix of their school's good and bad teachers. Overall, I find that the allocation of teacher and school inputs at the high school level contributes only 4% to the achievement gap between the top and bottom deciles of an index of student background. Finally, I find that schools that disproportionately serve disadvantaged students tend to be more sensitive to teacher quality. *JEL Codes:* I24, I21, J45, J62.

*I have greatly benefited from the input of Joseph Altonji, as well as Fabian Lange, Costas Meghir, Melissa Tartari, Lisa Kahn, and Amanda Kowalski. 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.

I. 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.

Furthermore, even if one finds that good teachers are concentrated at schools with well-supported students, policies that incentivize relocation of high quality teachers may not succeed in raising the performance of the most disadvantaged students. First, these teachers may get systematically assigned to the honors students within their new schools. Second, student performance at the targeted schools may be relatively insensitive to teacher quality. For example, some schools' facilities may be in such a state of disrepair or their disciplinary policies may be so lax that even good teachers cannot raise the performance of their students.

The bulk of 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? In the last section of the paper, I also provide suggestive evidence regarding a third question: (3) Would the performance of such disadvantaged students improve substantially if they were taught by better 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

1. See e.g. Rockoff (2004); Rivkin, Hanushek, and Kain (2005).

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 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 separately identify average teacher quality from other school-level inputs. They do test for 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, 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 (2010), whose focus is on teacher-school match effects, rather than the impact of teacher allocation on student performance inequality.

While one strand of the literature investigates the extent to which schools serving disadvantaged students hire teachers with inferior credentials,[5] and 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

2. See e.g. Hanushek *et al.* (2005).

3. See e.g. Aaronson, Barrow and Sander (2007).

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 (2008) does examine the collective effect of high 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 (2010); Rothstein (2010);

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 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 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, although teacher impacts may be smaller at the high school level (since teachers only spend about an hour a day with any given class), they 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 1997-2006. Such rich data permit identification and estimation via non-linear least squares of a flexible education production function that features both school- and teacher-specific intercepts as well as school-specific sensitivity parameters that scale the impact of a teacher's quality. Non-parametric identification of the joint distribution teacher quality and school quality stems from a large network of nearly 3,000 teacher transfers, coupled with testable exogenous mobility and conditional random assignment assumptions. Such tests fail to reject the assumption of exogenous mobility, and tests for conditional non-random assignment of students to teachers find no evidence of dynamic tracking, though they do suggest that teacher quality may not be perfectly persistent. Given estimates of each teacher's quality, I then characterize the way quality teaching is currently being allocated.

Finally, in the last section of the paper I address the third question by extending the model to include school-specific sensitivity parameters that scale the impact of a teacher's quality. This specification restricts a teacher's true teaching talent to remain one-dimensional (enabling a meaningful discussion about teacher quality), but allows his/her effective ability to raise test scores to vary across schools. I show that with a sufficiently rich transfer network the joint distribution of school quality, teacher quality, and school-specific sensitivity to teacher quality can be identified.

and Sass *et al.* (2010).

By allowing the impact of effective teaching to depend on the school, I can examine whether underprivileged students are in school or community environments that would enable them to profit from better teaching if it were provided. Allowing for complementarity between school and teacher quality also creates scope for efficiency gains from reallocating teachers.

The next three paragraphs summarize my answers to the three 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 .23 student test score standard deviations, enough to move an average student from the 50th test score percentile to the 59th percentile. 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 explained 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 .071 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 53rd percentile. Moreover, variation in teacher experience across schools, while substantial, 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 school's good and bad teachers across the courses they take, so that differences in quality among teachers from the same school only make a minor contribution 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 worse schools with worse teachers, but that the magnitudes of these differences are 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, only about 2% of the gap in performance between the top and bottom deciles of a regression index of student background can be attributed to differences in the school and teacher inputs these students receive at the high school level.

Third, I find that schools do seem to exhibit significantly different sensitivities to teacher quality, and interestingly, schools whose students are less academically prepared entering high school tend to be among the more sensitive schools. This implies that incentives to further equalize teacher quality would involve no efficiency-equality tradeoff. Simulations indicate that the optimal allo-

cation of the best teachers to the schools most sensitive to teacher quality would increase state average test scores by .09 standard deviations, decrease test score variance by about 5%, and increase the scores of students in the bottom 10% of a student background regression index by .17 standard deviations.

The remainder of the paper is structured as follows. Section II presents the educational achievement production function along with the assumptions required to justify its form. Section III discusses identification of the parameters of the function. Section IV describes the data. Section V presents the estimation strategy. Section VI presents the estimated distributions of teacher quality and school average teacher quality. Section VII 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 VIII examines the contribution to achievement inequality of the existing allocation of teachers and schools to students. Section IX tests the conditional random assignment assumption and examines out-of-sample predictive power. Section X generalizes the model to allow for school-specific sensitivity to teacher quality and presents evidence regarding the scope for efficiency gains from reallocating high quality teachers to struggling schools. Section XI interprets the findings and concludes.

II. The Education Production Function

Following Todd and Wolpin (2003), I assume that the test score performance performance of student i at time t, Y_{it}, is produced using a combination of current and past inputs of the student $\{\mathbf{A_i^t}\}$, family $\{\mathbf{F_i^t}\}$, peers $\{\mathbf{P_i^t}\}$, teachers $\{\mathbf{R_i^t}\}$, and schools $\{\mathbf{S_i^t}\}$, subject to measurement error e_{it}. To make the model tractable, I assume that current and past inputs are additively separable.[10] I also assume that current teacher and school inputs are additively separable from current student, family, and peer inputs, and from the measurement error.[11] Furthermore, since the North Carolina

10. This restricts the form of path dependence; while the effects of past inputs are allowed to persist, they do not affect the sensitivity of the student to his current inputs. For example, the model cannot capture the notion that some teachers only teach subject matter, while others teach children how to learn for themselves or pay attention (skills that may affect a student's ability to learn from his next teacher regardless of subject). One might also imagine that who a student's peers were in previous classes partly determined his current friends, which might affect his sensitivity to current inputs (particularly current peer inputs). This also cannot be captured by the model.

11. For example, this rules out the possibility that teachers have comparative advantages in working with certain kinds of students or parents. While this assumption may seem restrictive, most of the existing research suggests that specialization in teaching to particular parts of the student ability distribution is of secondary importance relative to vertical differences in teacher skill (Hanushek *et al.* [2005]; Lockwood and McCaffrey [2009]; Aucejo (2012) finds a somewhat larger role for teacher comparative advantages with respect to student background achievement). This

data I use provides scores from tests based on ten distinct high school subjects, several of which may be taken in the same school year, the education production function must reflect the fact that the mapping from inputs to test scores may be course-specific. Combining these features, we obtain:

$$(1) \qquad Y_{ict} = f^c(\mathbf{F_i^{t-1}}, \mathbf{A_i^{t-1}}, \mathbf{P_i^{t-1}}, \mathbf{R_i^{t-1}}, \mathbf{S_i^{t-1}}) + g^c(\mathbf{F_{it}}, \mathbf{A_{it}}, \mathbf{P_{it}}) + h^c(\mathbf{R_{it}}, \mathbf{S_{it}}) + e_{it}$$

where the c subscript in $f^c(*)$, $g^c(*)$, and $h^c(*)$ specifies the course in which the test being considered was taken. To minimize the impact of different choices of scales for exams taken in different subjects, I standardize each test score relative to the appropriate course-year-specific state distribution, and re-interpret Y_{ict} accordingly.

Let the school and teacher associated with student i in course c at year t be denoted by $s(i,t)$ and $r(i,c,t)$, respectively. I model the contribution of school and teacher inputs to student achievement as follows:

$$(2) \qquad h^c(\mathbf{R_{it}}, \mathbf{S_{it}}) = \delta_{s(i,t)c} + \phi_{s(i,t)t} + \mu_{r(i,c,t)} + d(ex_{r(i,c,t)t}) + \nu_{r(i,c,t)t}$$

The first component, $\delta_{s(i,t)c}$, 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 average 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 treated as a parameter to be estimated (one for each school-course combination).

The second 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. $\phi_{s(i,t)t}$ will be treated as an error component during estimation, and standard errors will be adjusted accordingly.

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 the third component, $\mu_{r(i,c,t)}$. Note that a teacher's ability to increase test scores is as-

assumption also does not permit the sensitivity of students' scores to teacher or school inputs to depend on their parents or their own aptitude. In other words, family, individual, and peer inputs are assumed to be substitutes for teacher and school inputs.

sumed to be common across subjects. This restriction is necessary in order to be able to pool 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.[12] To the extent that a teacher's quality is subject-specific (i.e. 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 the teacher teaches over the sample period, weighted by the frequency with which the teacher actually teaches those subjects. $\mu_{r(i,c,t)}$ is treated as a parameter to be estimated, and I will generally refer to $\mu_{r(i,c,t)}$ as teacher quality in the remainder of the paper.

The fourth component, 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. $d(ex_{r(i,c,t)t})$ will be flexibly parametrized using indicators for narrow ranges of experience and estimated.

Finally, idiosyncratic year-specific deviations in a teacher's performance from the path defined by his/her baseline ability and experience are captured by $\nu_{r(i,c,t)t}$. 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 teaches most effectively or intensively. $\nu_{r(i,c,t)t}$ will be treated as an error component during estimation.

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

Let $\mathbf{I_i}^{t-1} = \{\mathbf{F_i^{t-1}}, \mathbf{A_i^{t-1}}, \mathbf{P_i^{t-1}}, \mathbf{R_i^{t-1}}, \mathbf{S_i^{t-1}}\}$ denote the vector of past inputs, and let $\mathbf{I_i}^{t} = \{\mathbf{F_{it}}, \mathbf{A_{it}}, \mathbf{P_{it}}\}$ denote the vector of current family, individual, and peer inputs. We obtain:

$$Y_{ict} = f^c(\mathbf{I_i}^{t-1}) + g^c(\mathbf{I_i}^{t}) + \delta_{sc} + \phi_{st} + \mu_r + d(ex_{rt}) + \nu_{rt} + e_{ict} \tag{3}$$

12. 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 deems different subjects within a field to be interchangeable. In parallel work (Cook and Mansfield (2013)) I am testing the hypothesis that teachers have comparative advantages in particular subjects. Very preliminary results suggest that inherent comparative advantages are non-existent, but that subject-specific experience matters above and beyond total years of experience.

Even with the above assumptions, estimating the production function would still require one to observe all relevant current and prior student, family, and peer inputs. Thus, I instead narrow my focus to the estimation of the parameters most relevant to evaluating the contribution of high school teacher and school inputs to student inequality: the set of persistent teacher qualities $\{\mu_r\}$, the set of persistent school-course qualities $\{\delta_{sc}\}$, and the profile of teacher growth with experience $d(*)$. Given this focus, I use a vector of English and math test scores from 7th and 8th grade and their squares as a proxy for the impact of prior inputs on student test scores (which I denote $\tilde{\mathbf{Y}}_{\mathbf{i}}^{\mathbf{t-1}}$). 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.[13] The impact of using proxies for $f^c(\mathbf{I_i}^{t-1})$ and $g^c(\mathbf{I_i}^{t})$ is discussed further in the next section.

Thus, the specification I estimate is:

$$Y_{ict} = \tilde{\mathbf{Y}}_{\mathbf{i}}^{\mathbf{t-1}}\alpha_{\mathbf{c}} + \mathbf{X_{ict}}\beta_{\mathbf{c}} + \delta_{sc} + d(ex_{rt}) + \mu_r + \epsilon_{ict} \tag{4}$$

where the error term, ϵ_{ict}, is composed of:

$$\epsilon_{ict} = (f^c(\mathbf{I_i}^{t-1}) - \tilde{\mathbf{Y}}_{\mathbf{i}}^{\mathbf{t-1}}\alpha_{\mathbf{c}}) + (g^c(\mathbf{I_i}^{t}) - \mathbf{X_{ict}}\beta_{\mathbf{c}}) + \phi_{st} + \nu_{rt} + e_{ict} \tag{5}$$

13. 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.

III. 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 control function that eliminates bias from any persistent patterns of student sorting.

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 lower levels of schooling. 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 information 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

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

$$E[f^c(\mathbf{I_i}^{t-1}) + g^c(\mathbf{I_i}^t)|1(s(i,t) = s'), \tilde{\mathbf{Y}}_i^{t-1}, \mathbf{X}_{ict}] \; \forall \, r' \in \mathcal{R}, s' \in \mathcal{S} \tag{6}$$

At the high school level, crafting students' schedules is an onerous task done 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 still choose which difficulty levels to assign to which teachers, 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.

While the full set of observed inputs is laid out when I discuss the data in Section IV, 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. Since these measures are likely to serve as an effective proxy for the difficulty level of a class, conditional random assignment of students to teachers is likely a reasonable assumption at the high school level. Furthermore, note that non-random assignment of students to teachers, even conditional on student and peer observables, need not contaminate estimates of the between-school variance in teacher quality. Suppose, for example, that students are assigned to teachers in such a manner so that teacher quality is positively correlated with the unobserved component of their students' other inputs. Then the quality of a school's relatively effective teachers will be overestimated relative to the quality of the relatively ineffective teachers, and the estimated variance in teacher quality within schools will be biased upwards. However, note that the average bias in μ_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 taught by **some** teacher. Consequently, if transfer decisions are unrelated to the **bias** in transferring teachers' quality estimates,[14] then if there are enough transferrers connecting a school to the network, the average bias among transferring teachers will tend to zero, and the estimate of the average quality of the schools' teachers will be unbiased. 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 IX using the procedure introduced by Chetty et al. (2011) and Jackson (2012). While I do find that estimated teacher effects have less (though still considerable) predictive power out of sample than the model would suggest, I find no evidence that non-random assignment of students to teachers within courses is an important source of bias.

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

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

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.[15] 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 filling in gaps in, say, their child's physics knowledge. In this case, their only option would be to pay for costly professional tutoring. Nonetheless, in Section X, I generalize the production function to allow for the impact of teacher quality to be differentially important in different schools. This generalization creates scope for efficiency gains from reallocating teachers across schools. To the extent that such input compensation is common to students from the same school, it will be reflected in a smaller estimate of the school's sensitivity to teacher quality. While the estimates from the more general specification are noisier, its estimates of teacher quality, and by extension, estimates of the between-school variance in teacher quality, are robust to variation in the extent of input compensation across schools.

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 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.[16]

15. 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.

16. 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.

A number of existing articles have already shown that assumptions like those just discussed above are enough to identify and consistently estimate the distribution of teacher quality within schools.[17] Comparing the quality of teaching across schools, however, requires extra assumptions and information. Kramarz, Machin, and Ouazad (2008) decompose test scores of English primary school students into school and student components using a two-way fixed effects specification. They prove that distributions of school and student fixed effects can each be identified up to scale if two conditions are met. The first is that schools and student transfers form a connected graph (with schools as vertices and transferring students as edges). The second is that a student's choice to transfer to a different school is not correlated with changes in unobserved inputs. 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, identification requires that (1) 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, and (2) a teacher's choice to transfer to a different school or course is not correlated with any error component (i.e. it does not predict the level of their student's unobservable inputs or deviations in school or teacher performance from their long-run averages).

The existence of a large number of teachers transfers across schools, combined with a substantial fraction of teachers who teach multiple courses or switch courses during the sample, ensures that the sufficient mobility condition is easily satisfied for a large network of teachers and school-course combinations in the data. Discussion of the connectedness of the network of school-course combinations takes place in Section IV. As I discuss shortly, the extent of teacher mobility is important for the precision of parameter estimates.

We can state the exogenous mobility assumption more formally as:

Assumption 2: Exogenous Mobility

$$E[\epsilon_{ict}|s(i,t) = s', c = c', r(i,c,t) = r';\ r' \text{ transferred to combination } (s', c') \text{ at some } t' <= t] = 0 \quad (7)$$

$$E[\epsilon_{ict}|s(i,t) = s', c = c', r(i,c,t) = r';\ r' \text{ transferred from combination } (s', c') \text{ at some } t' > t] = 0$$

17. e.g. Hanushek *et al.* (2005); Kane and Staiger (2008)

where ϵ_{ict} is the composite error term defined in (5).

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 his/her course at his/her school by comparing the average residuals of his/her students' test scores with those of the other teachers, after removing the predicted impact of student- and classroom-level inputs.[18] If a teacher has taught at multiple schools, then he/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 distribution of within-school teacher quality in both subjects. With only one teacher linking schools (subjects), we would need to assume that his/her ability to increase test scores is the same across the two schools (subjects) in order to infer relative school (subject) quality from his/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 do not **systematically** perform better 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, though, 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 (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 (5) for ϵ_{ict} in (7), we observe that a systematic relationship

18. 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.

between a teacher's transfer decision and any of these components would violate Assumption 3. In Supplementary Section VI, I discuss an array of 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.

The second mechanism, related to the findings of Jackson (2010), is that teachers may systematically transfer toward schools or courses where they are idiosyncratically well matched. The model presented in Section II 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 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, or a school may only have a teacher switch from Algebra 1 to Algebra 2 if they suspect the teacher will have a comparative advantage at teaching Algebra 2. 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 Supplementary Section VI for further intuition). I find that such targeted mobility is statistically insignificant, and not of sufficient magnitude to substantially bias estimates in schools that do exhibit positive or negative net flows of teacher transfers.

IV. DATA

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 1997-2006 in up to ten subjects, along with a host of student, teacher, and school characteristics. During the sample period, North Carolina offered a standard curriculum with mandatory end-of-course tests for the

following subjects: Biology, English 1, U.S. History, Econ/Law/Politics, and Algebra 1, Algebra 2, Geometry, Physics, Physical Science, and Chemistry.[19]

The data contain a large number of observable current student[20] and peer[21] 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: first year teacher, second year teacher, third year teacher, years 4-6, years 7-12, and more than 12 years of experience. $d(x)$ is assumed to equal $d(x')$ for x, x' in the same experience cell[22]

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

19. 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. US History was not tested between 2004 and 2005.

20. Observable student inputs include the student's race and gender, indicators for parents' education categories, indicators for learning disabilities in writing, math, and reading, limited English proficiency, whether the student is gifted in math or English, and indicators for 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. I also allow for race matching effects between teacher and student, in acknowledgement of the findings of Hanushek *et al.* (2005). This is the one exception to the general rule that the sensitivity of student performance to student characteristics is not allowed to depend on teacher characteristics.

21. Observable peer/classroom inputs consist of class size, 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.

22. I limited the experience profile to six cells to permit comparability with the estimates of the non-linear specification introduced in Section X. Constructing the design matrix for the school-teacher-experience cell combinations used in the first stage of estimation for the non-linear specification (see Supplementary Section III) is computationally intensive, and increases rapidly in the number of experience cells.

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. Supplementary Section II describes the implementation of the fuzzy matching algorithm in detail. The dataset contains approximately 6 million test scores from over a million students, with 22,000 teachers, in 375 public high schools (with 3,750 potential school-subject combinations). Recall that identification requires a connected graph of schools 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 means that there are nearly 3,000 teachers who transfer across schools, and teachers have often taught hundreds of students. To be included in the sample, I required teachers to have taught at least 20 students. When I impose the restriction that the network of schools be 2-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] 329 schools and 3,091 school-subject combinations remain. In fact, the majority of the 329 schools are quite well connected: 217 of them are connected to each other by at least 10 transfers. Figure Ia shows the distribution across schools of the number of connecting transfers. Figure Ib shows the distribution of exams administered across teachers in the sample. Figure Ic 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.[25] Figure I presents a transition matrix describing the consider mobility of teachers across subjects. The first number in the (i,j)-th cell provides the number of teachers observed teaching both the row and column subjects at some point in their careers (not necessarily at the same time). The second number provides the fraction of teachers ever observed teaching the row subject that are also observed teaching the column subject in some year. The figure reveals that a large fraction of teachers teach multiple courses within their license (i.e. 2-3 science classes, or 2-3 math classes),

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.

25. About half of the teacher 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. Unfortunately, the data do not indicate whether a transfer was voluntary or not. Transfers across districts, like most employer changes, require the mutual agreement of the teacher and the new district.

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 taught at unconnected schools, and test scores from classes that were unmatchable, the data still contain 4,016,343 test scores from 855,238 students and 18,498 teachers. The test scores are scaled scores that I re-standardized to have zero mean and unit variance within each subject-year combination.[26]

V. ESTIMATION

From Section II, the model to be estimated is:

$$Y_{ict} = \tilde{\mathbf{Y}}_{\mathbf{i}}^{\mathbf{t-1}} \alpha_{\mathbf{c}} + \mathbf{X_{ict}} \beta_{\mathbf{c}} + \delta_{sc} + d(ex_{rt}) + \mu_r + \epsilon_{ict} \tag{8}$$

The standard estimation technique in the literature is the Empirical Bayes method advocated by McCaffrey *et al.* (2009) and Kane and Staiger (2008). By treating school and teacher quality as random effects, this technique only requires the estimation of the underlying joint distribution of teacher and school quality, and imposes multivariate normality to construct estimates of the posterior mean belief regarding the quality of each individual teacher and school. However, calculation of standard errors of the estimates of average teacher quality among the specific set of teachers that a particular student experienced (necessary for the analysis below) requires knowledge of the covariance matrix of estimation errors among the particular members of this set. Similarly, testing

26. Meghir and Rivkin (2010) have noted that if monotonic transformations of test scores convey the same information about student learning, a particular form of the education production function may be specific to an arbitrarily chosen test score scale. This is of potential concern in this context, since the method relies upon pooling tests from different subjects and years. However, most widely-used standardized tests have undergone considerable field-testing and analysis using item response theory to ensure that the final test instrument properly evaluates mastery of the relevant content and effectively differentiates between students throughout the distribution of learning. Consequently, assuming that all such tests, once standardized, share a common relationship with teacher and school inputs is not unreasonable. Thus, one way to interpret the estimates is that they measure the impacts of teachers or schools on standardized z-scores from field-tested exams designed for the whole range of student abilities, and would be general to any test fitting this description. Applying convex and concave monotonic transformations of test scores and re-estimating has little impact on the results below, except in cases where such transformations introduce floor or ceiling effects. Supplementary Figures 1-3, which plot histograms for the tests associated with each subject-year, show that the tests used in this analysis are not plagued by floor effects nor ceiling effects. See Koedel (2010) for an analysis of the impact of floor and ceiling effects on value-added estimates.

whether student performance in high-poverty schools is as sensitive to teacher quality as student performance in low-poverty schools requires knowledge of the full covariance matrix of sampling errors involving these schools, since sampling error will generally be correlated across schools.

Furthermore, a key disadvantage of the two-step residual method used by many papers in the literature is that any covariance between student observable characteristics and true teacher quality driven by 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.

Thus, I instead estimate equation (8) in one step via ordinary least squares. While there are over 4 million test score observations and nearly twenty thousand parameters, the extreme sparsity of the design matrices makes computation feasible. Since 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, 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. The teacher effects ($\hat{\mu}$) then adjust to fit the between-teacher variation that remains after removing the predicted regression index of student performance, $X_i\hat{\beta} + \tilde{Y}_i\hat{\alpha}$. Since the student-level coefficients are only identified using within-teacher variation, they will not reflect any covariance between teacher quality and average student quality. 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.

Analytical asymptotic standard errors are calculated for all parameters. In order to make estimation of the variance-covariance matrix computationally feasible, the calculation needed to be 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. Supplementary Section IV 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 $\overline{\mu}_s$. Supplementary Section V describes the procedure in detail.

VI. RESULTS

VI.A. Variance Decomposition

At first glance, differences in average teacher quality and teacher experience across schools seem to have the potential to create considerable disparities in test scores between schools. Table II shows that average teacher credentials differ substantially across schools. A school at the 5th percentile of the average teacher experience distribution has teachers with an average of eight fewer years of experience than a school at the 95th percentile. For fraction of teachers with Master's degrees, the 95th-5th quantile difference is .34. Also, substantial performance disparities across schools do exist that require explanation. Table III, which contains a decomposition of the variance in student test scores into within-school and between-school components, indicates that 8.7% 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 the 95th percentile is nearly a full student-level standard deviation (32nd percentile vs. 68th percentile of the test score distribution).

A closer look at Table III reveals that average teacher quality (and even school quality) in fact have very limited scope to explain average test score differences across schools. Rows 2 and 5 show that the lion's share (90.5%) 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 nearly three-quarters of the between-school variance. Row 3 indicates that unexplained variation between school-teacher-experience cells accounts for 5 percent of the total variance, suggesting that there is still scope for teacher quality to matter. However, Row 7, labeled "Total School Quality", shows that only 0.9 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 shockingly 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 $\tilde{\alpha}$. 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. Even though 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 .31 student-level standard deviations, enough to move an average student from the 44th percentile to the 56th percentile.

VI.B. Teacher Experience

The estimated values of the teacher experience profile, $d(\hat{ex})$, are presented in Table IV. First-year teachers are .059 student-level standard deviations worse than second-year teachers, .091 worse than third year teachers, .121 worse than teachers in their fourth through sixth years, .141 worse than teachers in their seventh through twelfth years, and .152 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 displayed in Row 4 of Table II, give the false impression that variation in average teacher experience across schools 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 IV displays various quantiles of the distribution of $\overline{d(ex)}_s$. The standard deviation is just .009, 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 .026 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.

VI.C. The True Variance of Teacher Quality, School Quality, and School Sensitivity to Teacher Quality

The first three columns of Table V display the raw and true variances of teacher quality $\hat{\mu}_r$ and school average teacher quality $\hat{\mu}_s$ for the baseline specification. While the standard deviation of $\hat{\mu}_r$ is .258, correcting for sampling error leaves an estimate of the true standard deviation in teacher quality of .229 student-level standard deviations. This estimate weighs 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. An average student who receives a teacher at the 75th percentile of teacher quality can expect to move from the 50th percentile to the 56th percentile, while one who receives a teacher at the 95th percentile can expect to move up to the 65th percentile, assuming test scores are distributed normally.[27] This is substantial, and generally in line with most estimates found in the literature at the elementary and middle school level.[28]

I calculate average teacher quality, denoted $\hat{\overline{\mu}}_s$, by weighting each teacher by 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 .071.[29] 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 52th (55th) percentile of the state test score distribution. So while average teacher quality does not vary dramatically across schools, attending a school with terrible teachers can still put a student at a meaningful disadvantage. Clearly, though, eliminating differences in average teacher quality across schools would not come close to eliminating test score gaps across schools. Figures II and III display the histograms of estimates of teacher quality ($\hat{\mu}_r$) and school average teacher quality ($\hat{\overline{\mu}}_s$), respectively, along with the underlying true densities of μ_r and $\overline{\mu}_s$, assuming a normal distribution.

Table VI presents 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 .23 in Algebra 1, .21 in Algebra 2, .20

27. This assumption is borne out by plots of the data. See Supplementary Figures 1-3

28. 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.

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

in Geometry, .22 in Biology, .26 in Chemistry, .22 in Physical Science, .25 in Physics, .24 in Econ/Law/Politics, .27 in U.S. History, and .21 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. Much of the between-school variation in teacher quality within each subject seems to idiosyncratic rather than common to all subjects at the school.

VII. 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 get consistently 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 systematically if students are choosing course tracks and the best teachers within a school tend to be assigned to the honors track.[30] On the other hand, if each student gets 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 five different courses, relative to the overall average teacher quality at the student's school. I denote this measure by $\hat{\overline{\mu}}_i$.[31] Using the delta method to calculate standard errors for each student's average teacher quality, σ_i^{μ}, I can estimate the variance in student-level teacher quality within schools as:

$$Var(\tilde{\overline{\mu}}_i) = Var(\overline{\mu}_i - \overline{\mu}_s) = Var(\hat{\overline{\mu}}_i - \hat{\overline{\mu}}_i) - (1/I)\sum_i (\sigma^{\tilde{\overline{\mu}}_i})^2. \tag{9}$$

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 qual-

30. 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.

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

ity of .077. 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 test scores in each course reduced (increased) by an average of .01 standard deviations, solely by virtue of the teachers he was assigned at his school. This is enough to move an average student from the 50th to the 46th (54rd) test score percentile. Thus, assignment of teachers to students within schools contributes about 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.[32]

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 .075 test score standard deviations.[33] Thus, within-school variation in the average teacher quality experienced across students does contribute to performance heterogeneity, but the variance in average

32. Note that subject-specific means were subtracted from estimated school-teacher-experience cell effects prior to the second stage of estimation, so that the average teacher **in each subject** has an estimated quality of zero. While it is possible that the average teacher in one subject may have, on average, better quality teachers than another subject, re-scaling test scores in each subject-year to have zero mean and unit variance precluded the examination of this possibility. Thus, "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.

33. 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, since 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.

teaching quality we observe is essentially equal to we would expect under random assignment.

VIII. The Impact of School and Teacher Inputs on Achievement Inequality

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

VIII.A. Are Underprivileged Students Systematically Being Assigned to Classes with Inferior Teachers?

One way to interpret the results from Section VII 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 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.

Thus, I also look more directly at whether students with observable characteristics that predict lower performance tend to receive their school's 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.[34] 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

34. 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.

their schools. I find that such students received teachers that were only .017 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 .019 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 .017 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 .010 above the mean among feasible paths using this restriction, while black students receive teachers who are .015 worse than the mean among feasible paths. Thus, I find only meager evidence that disadvantaged students are systematically being assigned relatively ineffective teachers at their schools, and to the extent that such evidence does exist, it explains a tiny fraction of the performance gaps we observe.

VIII.B. What Kinds of Schools Disproportionately Serve Underprivileged Students?

The estimated parameter distributions displayed in Section VI allow us to examine whether the schools disproportionately serving underprivileged students are actually the worst schools. To this end, Table VII 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 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 (Columns 3 and 4). 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 only .027 stu-

dent level standard deviations below average, while those in the top quartile have teachers who are only .023 standard deviations above average.

VIII.C. The Aggregate Contribution of School and Teacher Inputs to Inequality of Opportunity

While Table VII 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 school and 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 VIII displays the average values of school-average teacher quality and effective teacher experience and within-school deviations in teacher quality and effective experience from the school mean for each subpopulation.

I find that students in the bottom decile of the student background index attend schools where teachers are on average .013 test score standard deviations below the state mean. They are also assigned to classrooms with teachers that are on balance .011 test score standard deviations below their schools' averages. Below average teacher experience among their teachers reduces their test scores by an additional .005 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 .030 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 .015 above the state mean. They are assigned teachers that are .007 standard deviations above their schools' averages. Once teacher experience has been factored in, the teaching environment they enjoy raises their test scores by .027 standard deviations. However, the cumulative difference in the teacher inputs the two groups receive can only account for .057 of the 2.36 standard deviation difference in their test scores (2.4%). 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,

but the overall high school teaching environment they face lowers their test scores by only .008 standard deviations relative to the average student. However, given that their test scores are .513 standard deviations below average, it is clear that while teacher allocation in North Carolina high schools is not helping to fight racial achievement inequality, it is certainly not a major source of the problem. Differences in typical high school teaching environments only account for 2.3% of the black-white test score gap.

IX. 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,[35] 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 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 since every student taking a particular subject in a given year must be taught by some teacher, non-random sorting among teachers within a subject produces biases that sum to zero within subject. Consequently, bias from non-random student sorting can be eliminated by aggregating to the subject-year level. Thus, 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 1997-2006 form the original estimation sample, I employ data from the years 2007-2009 to construct estimates of student-weighted average teacher quality for each school-course-year combination, $\overline{\hat{\mu}}_{sct} = \frac{1}{N_{sct}} \sum_{r \in sc} w_r \hat{\mu}_r$, as well as estimates of average student test score residuals for each school-course-year combination: $\overline{Z}_{sct} = \frac{1}{N_{sct}} \sum_{i \in sct} Y_i - X_i\hat{\beta} - \tilde{Y}_i\hat{\alpha} - \tilde{\hat{\delta}}_{sc} - \hat{d}(exp(r(i)))$. As Chetty et al. (2011) point out, a univariate regression of the form

$$\overline{Z}_{sct} = \psi_0 + \psi_1 \overline{\hat{\mu}}_{sct} + \xi_{sct} \tag{10}$$

35. As discussed in Section III, 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 that are unobservably likely to declines. 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 Section VI.B of the Supplementary Material.

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. 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 (10) I shrink the $\overline{\hat{\mu}}_{sct}$ estimates toward zero to form empirical Bayes estimates of $\overline{\hat{\mu}}_{sct}$ via $\mu^{EB}_{sct} = \frac{\hat{Var}(\overline{\mu}_{sct})}{Var(\overline{\hat{\mu}}_{sct})}\overline{\hat{\mu}}_{sct}$[36]. 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 the form:

$$\overline{Z}_{sct} = \psi_0 + \psi_1 \mu^{EB}_{sct} + \xi_{sct} \tag{11}$$

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

Note that in order to form μ^{EB}_{sct} for a given school-course-year combination in equation (11), I need existing teacher quality estimates $\hat{\mu}_r$ from the 1997-2007 period for each teacher that taught in that school-course-year. New teachers cannot be left out, since they 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 2007-2009 that feature previously unobserved teachers leaves 1160 school-course-year observations.[37]

Column 1 of Table IX displays the results of estimating equation (11). The point estimate of $\hat{\psi}_1$ is .37, with a standard error of .06. While the aggregated teacher quality estimates clearly have considerable predictive power out-of-sample, a t-test easily rejects equality with 1. This contrasts sharply with the findings of Chetty (2011), who cannot reject equality with 1 in their preferred specifications. One possible explanation for a coefficient below 1 is that much of what the original μ_{hat} estimates capture is dynamic sorting of students unobservably likely to improve to particular teachers. However, an equally plausible alternative explanation is that teacher quality is not fully persistent. While I have included a common experience profile in the baseline specification, there may be a degree of heterogeneity in rates of improvement or decline among teachers. Furthermore,

36. 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 number of students between 1997 and 2006, 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 Supplementary Section V.B

37. 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.

changes in course or difficulty level assignments could require adjustments that are more difficult than most elementary or middle school teachers have to make, which would also serve to reduce the predictive power of teacher quality estimates derived from student performance from up to a decade earlier.

To disentangle these two explanations, 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:

$$\tilde{Z}_{rsct} = \frac{1}{N_{rsct}} \sum_{i \in sct} Y_i - X_i\hat{\beta} - \tilde{Y}_i\hat{\alpha} - \hat{\delta}_{sc} - \hat{d}(exp(r(i))) - \overline{Z}_{sct} \tag{12}$$

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:

$$\tilde{\hat{\mu}}_r^{sct} = \hat{\mu}_r - \sum_{r' \in \mathcal{SCT}} w_{r'}\hat{\mu}_{r'} \tag{13}$$

Consider regressing $\tilde{Z}_{rsct}$ on the appropriately shrunken empirical Bayes estimate[38] of $\tilde{\hat{\mu}}_r^{sct}$, denoted $\tilde{\mu}_{rsct}^{EB}$:

$$\tilde{Z}_{rsct} = \zeta_0 + \zeta_1\tilde{\mu}_{rsct}^{EB} + \tilde{\xi}_{rsct} \tag{14}$$

If the differences in performance among teachers from the same school-course-year are driven by a combination a 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 $\zeta_1 \rightarrow_p 1$. Note that this specification requires at least two teachers to be observed teaching in any given school-course-year cell. In order to avoid dropping teachers from school-course-years that feature a previously unobserved teacher, I actually implement (14) by calculating deviations in both student residuals and estimated teacher quality relative to a normalized teacher

38. See Supplementary Section V.B for a description of the shrinkage procedure for this specification

for whom a previous estimate of quality exists, which yields 1207 observations. Column 2 of Table IX displays the results of estimating equation (14). The point estimate of ζ_1 is .40, with a standard error of .07. These results suggest that dynamic tracking is not a major source of bias. However, teacher quality may not be as persistent across changes time or across changes in course or track assignments as it is at the elementary and middle school levels (where such tracks generally do not exist).

X. SENSITIVITY TO TEACHER QUALITY

This section provides suggestive evidence addressing the third question raised in the introduction: Would the performance of such disadvantaged students improve substantially if they were taught by better teachers? While the previous sections showed that the current teachers at the schools such students attend are not far below the state average, the substantial within-school variance in teacher quality suggests that compensatory policies that incentivize existing high quality teachers to transfer to underperforming schools may have the potential to considerably decrease performance inequality. However, the scope for potential gains from recruiting high quality teachers to staff underperforming schools depends on whether high quality teachers are able to be as effective in such difficult school environments.

Although the linear model employed to this point allows an effective characterization and decomposition of disparities in teaching inputs across more and less privileged schools and individuals, it imposes that teacher quality and school-level (and school-course-level) inputs are perfect substitutes. While most of the existing literature has also ignored the possibility that teacher quality may be complementary to other school-level inputs,[39] a variety of factors could affect the extent to which student performance responds to effective or ineffective instruction in a given school. In schools where after-school jobs are rare, students may have more time to do well-crafted homework assignments by good teachers, magnifying the impact of the teacher. Or, students may compensate for a bad teacher by spending more time reading the textbook, as highlighted in Section 3. Alternatively, poorly equipped classrooms may constrain the set of teaching tools available to teachers. Strict enforcement of common curriculum and assignments across different teachers' classrooms at a school may limit the damage done by an otherwise unprepared teacher, but mute the impact of

39. A few papers do consider possible complementarities between observable teacher characteristics and student-level inputs, such as race-matching effects (Hanushek *et al.* [2005]), and differential effects of teacher experience on students with different levels of past performance (Clotfelter, Ladd and Vigdor [2006]).

a particularly creative teacher. To capture this intuition, I introduce school-specific scaling factors, $\{\gamma_s\}$, that capture the extent to which a school's students respond to the quality of their teachers. The final empirical specification becomes: I generalize the specification of the contribution of school and teacher inputs as follows:

$$Y_{ict} = \tilde{\mathbf{Y}}_{\mathbf{i}}^{\mathbf{t-1}}\alpha_{\mathbf{c}} + \mathbf{X_{ict}}\beta_{\mathbf{c}} + \delta_s + \gamma_s[d(ex_{rt}) + \mu_r] + \epsilon_{ict} \tag{15}$$

γ measures the extent to which a school offers a learning environment that either amplifies or mutes the impact of a teacher on her students' performance. Note that many underlying school-level factors may contribute to both δ_{sc} and γ_s. For example, an overly lenient discipline policy might decrease both δ_{sc} and γ_s. However, since this analysis is focused on the impact on achievement of the overall school experiences received by different students, I do not model the consequences of specific school policies, opting instead for a flexible specification that places minimal restrictions on how a school can influence student performance. I treat each γ_s as a parameter to be estimated, and I will refer to it as the sensitivity of school s to teacher quality. I impose that such sensitivity is common to all courses within a school, since school-course specific parameters are either not identified or identified but incredibly noisy, depending on the frequency with which teachers who teach particular courses at a school transfer in our out of the school.

$d(*)$, $\{\mu_r\}$, $\{\gamma_s\}$, and $\{\delta_s\}$ are only identified up to scale. Hence, I normalize the experience profile so that $d(0) = 0$. I normalize γ_s so that the median across schools is one; each school's value of γ_s captures the school's sensitivity as a fraction of the sensitivity at the median school. μ_r is normalized so that it captures teacher r's ability to increase test scores relative to the average teacher in the sample at a school of median sensitivity. δ_{sc} is normalized so that it captures the expected increase in test scores from enrolling in course c in school s relative to a school-course combination with average δ. Together, the expression $\gamma_{s(i,t)}(\mu_{r(i,t)} + d(ex_{r(i,t)t}) + \nu_{r(i,t)t})$ can be interpreted as the "effective" quality of teaching that student i received in year t, as measured by the actual impact the teacher had on the student's test score.

There is also an empirical motivation for the generalized specification. Using elementary school data in Florida and North Carolina, Sass *et al.* (2010) find higher within-school variance in teacher fixed effect estimates among schools serving high poverty populations relative to those serving low poverty populations, and conclude that there is greater heterogeneity in teacher quality among high

poverty schools.

I also find evidence of excess variation in teacher-specific average residuals across schools in the high school context. Specifically, I first calculate average test score residuals (removing the effects of observable student characteristics) for each school-teacher combination, and take the variance of these average test score residuals at each school. I then compare the distribution of these school-specific variances across schools to the distribution that would occur naturally under random assignment of teachers to schools (holding the number of teachers at each school fixed). The actual variance in within-school variances of teacher residuals across schools is 2.5 times as large as that produced by a random allocation of teachers to schools. Further, I also find that such within-school variance in average residuals across teachers is between 10 and 15 percent larger in schools whose fraction eligible for free lunch eligibility and fraction minority place them in the top quartile of high schools in North Carolina.

However, in a linear value-added specification, one cannot tell whether high poverty schools are actually attracting teachers whose quality is more heterogeneous, or instead whether the same variance in underlying teacher skill is magnified in importance in a high poverty setting. Distinguishing between these two hypotheses is critical for determining whether such schools can be more efficiently improved by policies that alter the school environment versus policies that facilitate more careful screening of potential teachers. A specification incorporating both school-specific scaling factors as well as persistent teacher quality allows one to separately identify the contributions of actual heterogeneity in teacher skill and sensitivity to teacher quality to the excess variation in school-specific teacher fixed effect variances.

Surprisingly, identification of the non-linear model in equation (4), which includes school-specific sensitivity to teacher quality, does not require much more information. As long as teachers tend to naturally increase their quality as they gain experience, we have an extra source of within-teacher variation that informs us about school sensitivity to teacher quality. Thus, the set of school sensitivity parameters $\{\gamma^s\}$ can be identified by variation across schools in how quickly test scores increase with teacher experience, before even considering teachers who transfer. In practice, when there are more transferrers than the minimum required to construct one connected graph, the γ parameters will also be identified from variation across schools in the performance of higher quality transferrers relative to lower quality transferrers, independently of experience. For example, if two teachers both transfer from the same school A to a common second school B, and the difference between the two teachers' average test score residuals is larger at school B than at school A,

this will contribute to a relatively larger estimated γ value at school B. Thus, with school-specific sensitivity to teacher quality, the expected difference in student performance among a set of transferring teachers at two different schools will now reflect differences in sensitivity to the quality of these transferring teachers in addition to differences in school quality. A formal proof of identification, given the assumptions and conditions above, is provided in Supplementary Section I. The proof makes clear that identification of the quality of non-transferring teachers fails at schools that are perfectly insensitive to teacher quality. Consequently, estimates of teacher quality become extremely unstable when estimated sensitivity to teacher quality is close to 0. Thus, I constrain each school to be positive and at least 1/4 as sensitive to teacher quality as the median school ($\gamma_s \geq .25$).[40]

Figure IV displays the histogram of estimates of school sensitivity to teacher quality, ($\hat{\gamma}_s$), as well as the underlying true density of γ_s, obtained by correcting the variance in the estimates for sampling error and assuming a log-normal distribution. Rows 4-6 of Table V provide the estimates of the raw and error-adjusted variance of the model parameters for the augmented specification. Allowing school-specific sensitivity parameters leads to a small increase in the overall variance in true persistent teacher quality, but a small decrease in variance of average teacher quality across schools. Furthermore, considerable variation in γ_s exists: a school whose sensitivity to teacher quality is at the 10th percentile of the true distribution of school sensitivity is expected to be about .46 times as sensitive to teacher quality as the median school, while one whose sensitivity is at the 90th percentile is expected to be about 2.15 times as sensitive as the median school. A likelihood ratio test overwhelming rejects the restriction that $\gamma_s = 1 \; \forall \; s$.

Columns 5 and 6 of Table VII reexamines whether the schools disproportionately serving underprivileged students have inferior teachers using the more general nonlinear specification. While the results are similar to the baseline model, they suggest that failing to account for differential sensitivity to teacher quality may cause us to underestimate the degree to which superior teachers are cluster at schools serving more advantaged students. Columns 7 and 8 replace average esti-

40. A negative sensitivity would imply that the teachers who reduce the test scores the most are the teachers of highest true quality. A small or negative estimated sensitivity can be obtained if the students transferring teachers who were relatively more effective at other schools perform less well than the students of the transferring teachers who were relatively less effective at other schools. At schools with small samples of transferring teachers, test score measurement error, teacher-year error components, and idiosyncratic school-teacher match components need not wash out. Relaxing this constraint does not substantially alter the shape of the sensitivity distribution, but a handful of schools do exhibit small estimated sensitivities, and a few exhibit negative estimated sensitivities. Such small and negative sensitivities generally lead to implausibly large teacher quality estimates for non-transferring teachers, revealing that such sensitivities cannot possibly be reflecting true insensitivity or negative sensitivity to quality. However, a few such outliers can skew the measurement error adjustment substantially, which motivates imposing the parameter constraints.

mated teacher quality ($\hat{\mu}_s$) with estimated school sensitivity to teacher quality ($\hat{\gamma}_s$). They reveal an interesting result: schools whose average student characteristics generally predict lower achievement tend to be more sensitive to teacher quality. Schools in the bottom quartile of 8th grade math scores and schools in the top quartile of percent free lunch eligible and percent black have median values of $\hat{\gamma}$ that are around 40% to 50% higher than the overall median school in the sample. Most tellingly, schools in the bottom quartile of the $X_{ict}\beta + \tilde{Y}_i\alpha$ index have median sensitivity of 1.64, while those in the top quartile have median sensitivity of .74. These results suggest the possibility of efficiency gains from reallocating high quality teachers to low performing schools. As a back-of-the-envelope calculation to quantify the potential magnitude of such gains, in Supplementary Section IX I examine the predicted impact of implementing the efficient allocation of teachers to schools, in which the most sensitive schools receive the best teachers. I find that the efficient allocation increases the mean test score by .187 student-level standard deviations. Furthermore, the average test score among students in the bottom 10% of a regression index of student background ($X_{it}\beta + \tilde{Y}_i\alpha$) is .106 test score standard deviations larger under the efficient allocation than under the status quo allocation. These counterfactual estimates suggest that while school-teacher complementarity is strong enough to be meaningful, it is certainly not strong enough to make efficient use of teacher quality a policy priority. However, one may be comforted that policies that attempt to reallocate teacher talent for the sake of educational equality (such as bonuses for effective teachers who teach in poorer school districts) would be likely to have the side effect of increasing average test scores.

XI. 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 high schools in North Carolina. Disadvantaged students, as indicated by low values of the predictive index $X_{it}\beta + \tilde{Y}_i\alpha$, do tend to be exposed to slightly inferior high school teachers. However, the contribution of this source of inequality to overall test score achievement gaps is nearly negligible, 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 Supplementary Section VIII), 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).[41]

Another explanation may lie in the fact that teachers are hired by districts rather than schools, so that for the half of transfers that are occurring within districts, transfer patterns may more closely reflect the preferences of district administrators rather than 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 do not explain 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 .23 student-level standard deviations at a school of median sensitivity, enough to move an average student from the 50th to the 59th percentile of the state test score distribution.

Given the sizable variation in teacher quality within schools, I explore the impact that variation in the average teacher quality experienced while in high school has on 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

41. 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 teacher's quit rates and location decisions do respond systematically to some school characteristics can be seen in Goldhaber, Gross and Player (2007) and Jackson (2009).

school, the variation in average teaching quality experienced across students is fully explained 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.

Finally, I find evidence of substantial differences across schools in sensitivity to teacher quality that can amplify or mute the effectiveness or ineffectiveness of teachers. Interestingly, the schools that are estimated to be most sensitive to teacher quality tend to be those serving students whose characteristics are usually associated with low performance. Thus, policies that incentivize effective teachers to join and to stay at struggling schools would be likely to increase average test scores statewide in addition to the clear equity payoff.

What might cause student performance to respond more strongly to the quality of teaching in schools serving disadvantaged populations? One possibility is that the ability to maintain control of the classroom is a central component of teaching skill, and that its importance is magnified in contexts where parents and the principal are providing less support. Furthermore, classroom control techniques are likely to be primarily learned with experience, which would be consistent with the results reported above. A second possibility is that wealthier and better prepared students (and their parents) have the time and resources to compensate more strongly for inadequate teachers by seeking outside tutoring or studying textbooks outside of class. Investigating the conditions under which teachers can make a bigger difference is a ripe area for future research.

A caveat merits mention. The distributions of teacher quality and school sensitivity characterized by this paper reflect the equilibrium that existed in North Carolina between 1997 and 2006. 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 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

Abowd, John, Robert Creecy, and Francis Kramarz (2002) 'Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data.' Working Paper

Arcidiacono, Peter, Gigi Foster, Natalie Goodpaster, and Josh Kinsler (2011) 'Estimating Spillovers with Panel Data, with an Application to the Classroom.' Working Paper

Aucejo, Esteban (2011) 'Assessing the Role of Student-Teacher Interactions.' Working Paper

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.' Unpublished Manuscript

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

___ (2010) 'Match Quality, Worker Productivity, and Worker Mobility.' NBER Working Paper 15990, National Bureau of Economic Research, Inc

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

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

McCaffrey, Daniel, Tim Sass, J.R. Lockwood, and Kata Mihaly (2009) 'The Intertemporal Variability of Teacher Effect Estimates.' *Education Finance and Policy* 4(4), 572–606

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

XII. TABLES AND FIGURES

TABLE I: TEACHER MOBILITY ACROSS COURSES

Subject		Subject									
		Algebra 1	Algebra 2	Biology	Chemistry	E/L/P/C	English	Geometry	Physical Sciences	Physics	U.S. History
Subject	Algebra 1	4700 1.0	2213 0.471	193 0.041	49 0.01	168 0.036	220 0.047	2052 0.437	182 0.039	60 0.013	122 0.026
	Algebra 2	2213 0.769	2878 1.0	59 0.021	25 0.009	44 0.015	47 0.016	1424 0.495	64 0.022	41 0.014	35 0.012
	Biology	193 0.061	59 0.019	3189 1.0	435 0.136	155 0.049	193 0.061	75 0.024	1250 0.392	155 0.049	119 0.037
	Chemistry	49 0.036	25 0.018	435 0.321	1357 1.0	14 0.01	17 0.013	22 0.016	742 0.547	311 0.229	15 0.011
	E/L/P/C	168 0.051	44 0.013	155 0.047	14 0.004	3276 1.0	312 0.095	55 0.017	137 0.042	5 0.002	1514 0.462
	English	220 0.05	47 0.011	193 0.044	17 0.004	312 0.071	4391 1.0	69 0.016	150 0.034	2 0	200 0.046
	Geometry	2052 0.717	1424 0.498	75 0.026	22 0.008	55 0.019	69 0.024	2860 1.0	63 0.022	36 0.013	39 0.014
	Physical Sciences	182 0.069	64 0.024	1250 0.473	742 0.281	137 0.052	150 0.057	63 0.024	2643 1.0	430 0.163	97 0.037
	Physics	60 0.09	41 0.061	155 0.232	311 0.465	5 0.007	2 0.003	36 0.054	430 0.643	669 1.0	6 0.009
	U.S. History	122 0.046	35 0.013	119 0.045	15 0.006	1514 0.572	200 0.076	39 0.015	97 0.037	6 0.002	2647 1.0

The top entry in the i,j-th cell is the number of teachers who are observed teaching in both the i-th and the j-th subject (not necessarily in the same year).
The bottom entry of the i,j-th cell is the fraction of teachers ever observed teaching the i-th subject who are also observed teaching the j-th subject at some point during the sample.

TABLE II: THE DISTRIBUTION OF TEACHER CREDENTIALS ACROSS SCHOOLS

Selected Quantiles of the Distribution of School Means of Selected Teacher Credentials					
Credential	5th	25th	50th	75th	95th
Fraction of Teachers w/Masters or Other Advanced Degree	.110	.196	.259	.332	.455
Fraction of Teachers w/ National Board Certification	0	.013	.039	.067	.138
Fraction of Teachers Who Are Uncertified	.053	.102	.138	.187	.284
Years of Teaching Experience	8.67	11.13	12.65	14.23	16.75

TABLE III: VARIANCE DECOMPOSITION OF STUDENT TEST SCORES

	Variance Component	Variance	Standard Deviation	Fraction of Total Var.
(1)	Total: $Var(Y_{ist})$	.903	.95	–
	Components:			
(2)	Student Background $Var(X_{ict}\beta_c + \tilde{Y}_i^{t-1})$	.505	.711	.559
(3)	Effective School and Teacher Quality $Var(\lambda_{srj})^*$	.052	.228	.058
(4)	Cov(Stu. Background, Eff. Sch./Tch. Qual.) $2 * Cov(X_{ict}\beta_c + \tilde{Y}_i^{t-1}, \lambda_{srj})$	.030	–	.034
(5)	Idiosyncratic Test Score Error $Var(\epsilon_{ict})$	.313	.560	.346
(6)	Between School Total: $Var(\overline{Y}_s)$	.079	.280	.087
	Components:			
(7)	School Average Student Background $Var(\overline{X_s\beta_c} + \overline{\tilde{Y}}_s^{t-1})$	.058	.240	.064
(8)	Total School Quality $Var(\overline{\lambda}_s)^{**}$	.010	.098	.011
(9)	Cov(Avg. Stu. Background, Total Sch. Qual.) $2 * Cov(\overline{X_s\beta_c} + \overline{\tilde{Y}}_s^{t-1}, \overline{\lambda}_s)$	.011	–	.012

$^*\lambda_{srcj}$ 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 (See Supplementary Section III). $\lambda_{srcj} = \lambda_h = \delta_{sc} + \mu_r + d(ex_{rt})) + \omega_{srcj}$. 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 IV: 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	.059	.091	.121	.141	.152
	(0)	(.004)	(.004)	(.004)	(.005)	(.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	-0.026	-0.015	-0.006	0.006	0.013	0.017
Test Score Percentile (Avg. Student)	49.0%	49.4%	49.8%	50.3%	50.5%	50.7%

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 V: RAW AND ERROR-ADJUSTED VARIANCES IN μ_r, $\overline{\mu}_s$ AND γ_s: BASELINE AND DIFFERENTIAL SENSITIVITY SPECIFICATIONS

	Baseline $\delta_{sc} + \mu_r + d(exp_r)$			Differential Sensitivity $\delta_{sc} + \gamma_s(\mu_r + d(exp_r))$		
Parameter	Raw Var.	True Var.	True Std.	Raw Var.	True Var.	True Std.
Teacher Quality (μ_r)	.067	.052	.229	.138	.071	.267
	(.002)	(.002)	(.005)	(.016)	(.009)	(.017)
School Average Teacher Quality ($\overline{\mu}_s$)	.008	.005	.071	.011	.003	.053
	(.001)	(.001)	(.005)	(.001)	(.001)	(.010)
School Sensitivity to Teacher Quality (γ_s)	–	–	–	1.05	.676	.822
	–	–	–	(.110)	(.087)	(.053)

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

TABLE VI: RAW AND ERROR-ADJUSTED VARIANCES IN μ_r AND $\overline{\mu}_s$ BY SUBJECT (BASELINE SPECIFICATION)

	Total			Between-School		
Parameter	Raw Var.	True Var.	True Std.	Raw Var.	True Var.	True Std.
Algebra 1	.071 (.003)	.052 (.003)	.228 (.006)	.025 (.002)	.018 (.002)	.132 (.009)
Algebra 2	.063 (.003)	.045 (.003)	.212 (.006)	.025 (.002)	.018 (.002)	.134 (.008)
Geometry	.059 (.002)	.040 (.002)	.199 (.005)	.025 (.002)	.018 (.002)	.135 (.009)
Biology	.070 (.003)	.048 (.003)	.218 (.006)	.029 (.003)	.020 (.003)	.142 (.009)
Chemistry	.087 (.005)	.065 (.005)	.256 (.010)	.048 (.011)	.037 (.011)	.193 (.027)
Physical Science	.068 (.003)	.050 (.003)	.224 (.006)	.032 (.003)	.023 (.003)	.150 (.009)
Physics	.072 (.007)	.064 (.007)	.253 (.014)	.052 (.006)	.040 (.005)	.200 (.014)
Civics/ELP	.083 (.005)	.057 (.005)	.240 (.009)	.038 (.005)	.027 (.005)	.165 (.015)
U.S. History	.087 (.006)	.070 (.005)	.265 (.010)	.049 (.010)	.037 (.009)	.192 (.024)
English 1	.071 (.003)	.044 (.003)	.209 (.007)	.025 (.004)	.014 (.003)	.117 (.013)

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. Unfortunately, they are likely to be underestimates, since the individual parameter estimates are held fixed across bootstrap samples, rather than re-estimating the model using each bootstrap sample. Re-estimating the model (along with calculating analytical standard errors for individual parameters) for each bootstrap sample was computationally infeasible.

TABLE VII: AVERAGE SCHOOLING INPUTS AMONG SCHOOLS IN THE TOP QUARTILE VERSUS BOTTOM QUARTILE OF VARIOUS STUDENT CHARACTERISTICS: BASELINE AND DIFFERENTIAL SENSITIVITY SPECIFICATIONS

			Specification					
			Baseline Model		Differential Sensitivity			
	Mean Student Characteristic		Mean Teacher Quality ($\hat{\bar{\mu}}_s$)		Mean Teacher Quality ($\hat{\bar{\mu}}_s$)		Mean Sens. to Tch. Qual. ($\hat{\gamma}_s$)	
	Bottom	Top	Bottom	Top	Bottom	Top	Bottom	Top
Mean 8th Grade Math Score	-.155	.548	-.025 (.007)	.030 (.007)	-.033 (.011)	.036 (.020)	1.45 (.437)	0.76 (.472)
Percent Black	.058	.608	.013 (.008)	-.029 (.007)	.015 (.018)	-.039 (.011)	0.78 (.406)	1.40 (.474)
Percent Hispanic	.010	.086	-.028 (.008)	.008 (.006)	-.018 (.012)	-.004 (.011)	1.18 (.451)	1.08 (.442)
Percent Eligible for Free Lunch	.139	.516	.005 (.005)	-.028 (.009)	.001 (.009)	-.041 (.013)	0.91 (.394)	1.48 (.397)
Stu. Backgr. Index ($X_i\hat{\beta} + \tilde{Y}_i^{t-1}\hat{\alpha}$)	-.382	.260	-.027 (.008)	.023 (.006)	-.040 (.012)	.023 (.016)	1.64 (.438)	0.74 (.464)

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.

Mean Sens. to Tch. Qual. is the average value of estimated sensitivity to teacher quality ($\hat{\gamma}_s$) 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 VIII: 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. $(\hat{\delta}_s)$	Between Sch. Tch. Exper. $(\hat{\gamma}_s)$	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%	-1.18	-.013	-.011	-.003	-.002	-.030
Bottom 25%	-.862	-.010	-.009	-.003	-.001	-.022
Top 25%	.935	.010	.006	.004	.000	.021
Top 10%	1.28	.015	.007	.006	.000	.027
Race						
White	.212	.001	.002	.007	.001	.009
Black	-.513	-.005	-.014	-.002	-.002	-.008

"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: $(\overline{\hat{\mu}}_i + \overline{\hat{d}(ex)}_i)$.

TABLE IX: OUT-OF-SAMPLE TESTS FOR PERSISTENCE OF TEACHER QUALITY AND DYNAMIC TRACKING USING DATA FROM 2007-2009

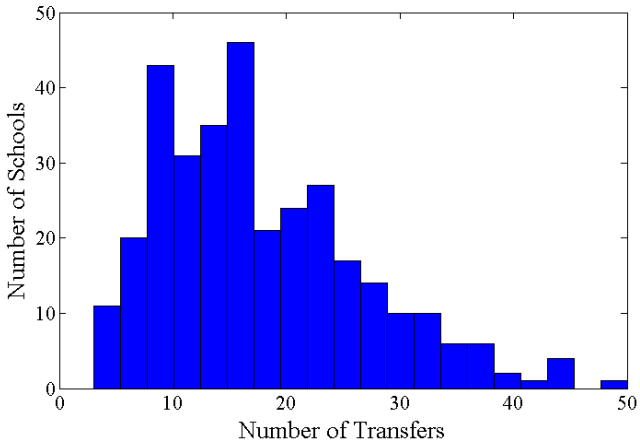
	Specification/Level of Observation	
	School-Course-Year Averages	Teacher-Specific Deviations from School-Course-Year Averages
	(1)	(2)
$\hat{\psi}_1$	.37 (.06)	–
$\hat{\zeta}_1$	–	.40 (.07)
N	1160	1207

"School-Course-Year Averages" refers to a specification in which school-course year averages of student residuals from the years 2007-2009 are regressed on school-course-year averages of predicted teacher quality (based on shrunken estimates from the years 1997-2006): $\overline{Z}_{sct} = \psi_0 + \psi_1 \mu^{EB}_{sct} + \xi_{sct}$

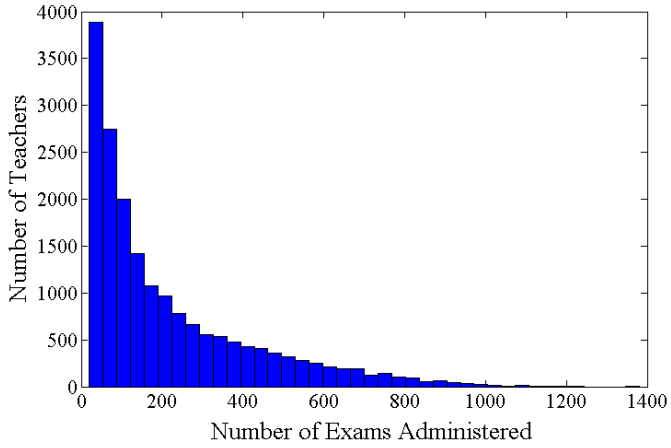
"Teacher-Specific Deviations from School-Course-Year Averages" refers to a specification in which teacher-school-course year specific deviations of averages student residuals from school-course-year means from the years 2007-2009 are regressed on teacher-school-course year specific deviations of predicted teacher quality from school-course-year means (based on shrunken estimates from the years 1997-2006): $\tilde{Z}_{rsct} = \zeta_0 + \zeta_1 \tilde{\mu}^{EB}_{rsct} + \tilde{\xi}_{rsct}$

FIGURE I: 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

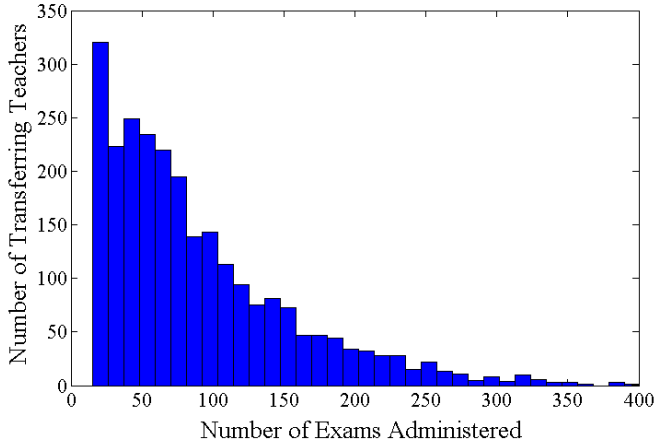


FIGURE II: INDIVIDUAL TEACHER QUALITY ESTIMATES ($\hat{\mu}_r$) AND THE UNDERLYING DENSITY OF TEACHER QUALITY (μ_r), AFTER CORRECTING FOR SAMPLING ERROR

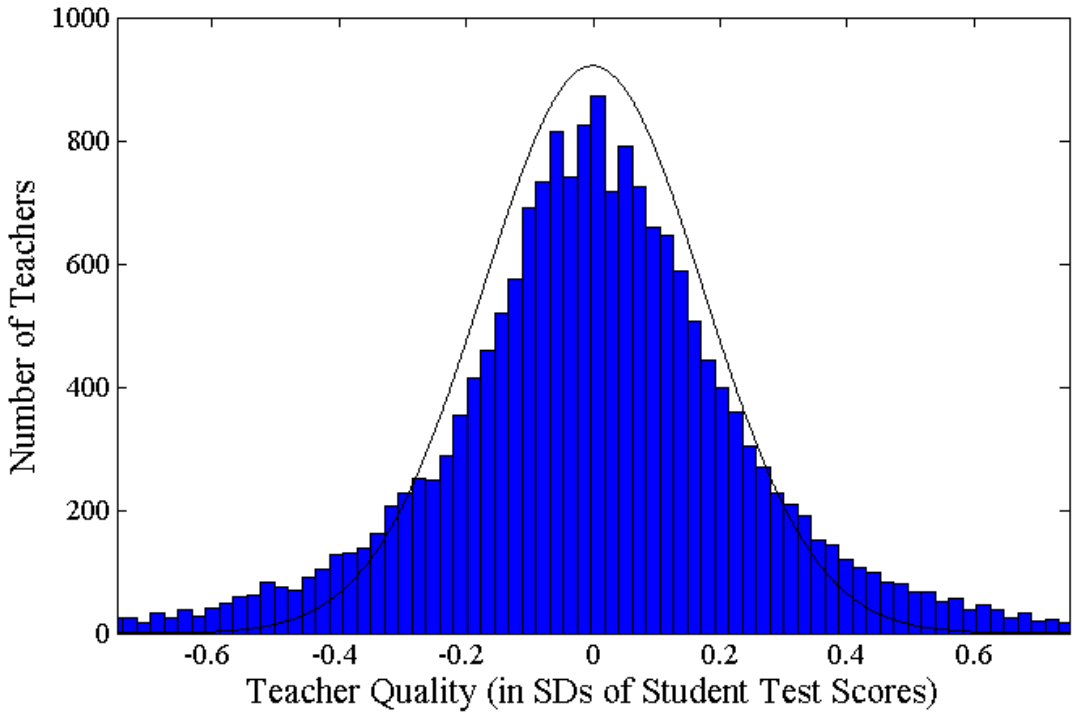


FIGURE III: SCHOOL AVERAGE TEACHER QUALITY ESTIMATES ($\hat{\bar{\mu}}_s$) AND THE UNDERLYING DENSITY OF SCHOOL QUALITY ($\bar{\mu}_s$), AFTER CORRECTING FOR SAMPLING ERROR

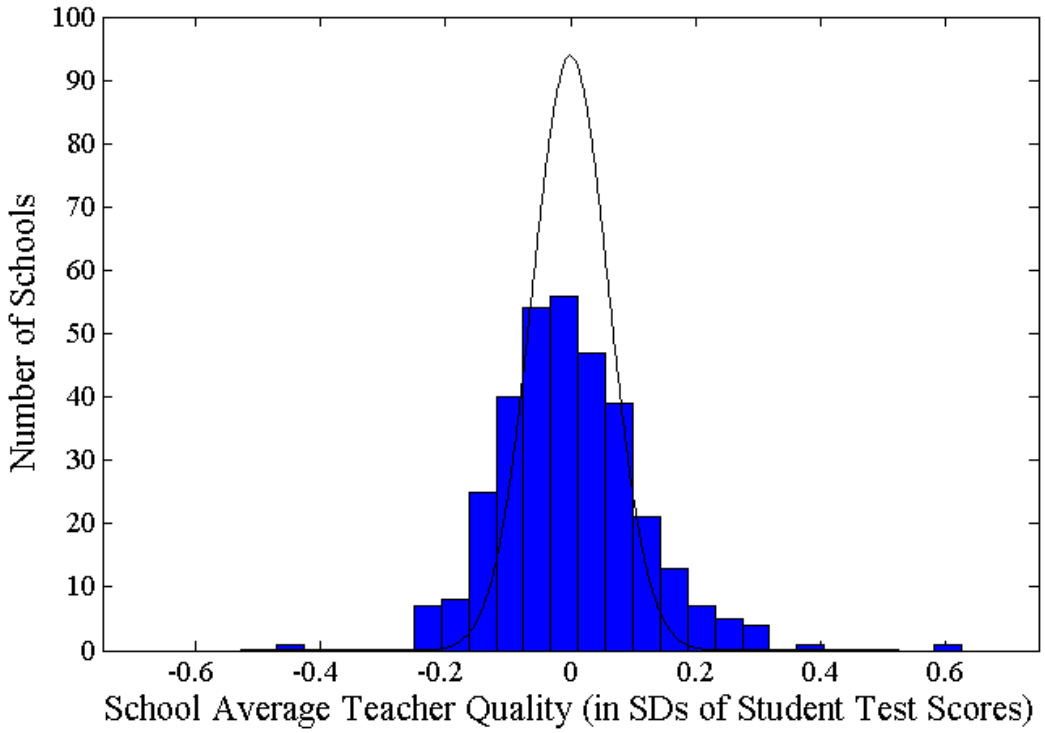


FIGURE IV: ESTIMATES OF SCHOOL SENSITIVITY TO TEACHER QUALITY ($\hat{\gamma}_s$) AND THE UNDERLYING DENSITY OF SCHOOL SENSITIVITY TO TEACHER QUALITY (γ_s), AFTER CORRECTING FOR SAMPLING ERROR

